Computer Modeling in Engineering & Sciences

Tech Science Press

https://doi.org/10.32604/cmes.2026.081475

ARTICLE

Jumper Line Detection Method for Situational Awareness of Aerial Lift Operations in Live-Line Maintenance of Overhead Distribution Systems

Joonhyeok Moon[1], Siheon Jeong[1], Byeonghyun Lee[1], Jeik Choi[1] and Ki-Yong Oh[1,2,*]

[1]Department of Mechanical Convergence Engineering, Hanyang University, 222 Wangsimni-ri, Seongdong-gu, Seoul, Republic of Korea

[2]School of Mechanical Engineering, Hanyang University, 222 Wangsimni-ri, Seongdong-gu, Seoul, Republic of Korea

*Corresponding Author: Ki-Yong Oh. Email: kiyongoh@hanyang.ac.kr

Received: 03 March 2026; Accepted: 18 May 2026; Published: 30 June 2026

ABSTRACT: Maintaining overhead distribution facilities inherently involves high risks for operators, where ensuring worker safety and operational efficiency remains a paramount challenge. In particular, automating the positioning of aerial work platforms is crucial to mitigate electrocution hazards during live-line maintenance tasks. This paper proposes a novel autonomous framework for detecting jumper lines that could be employed to estimate the optimal bucket position in live-line maintenance of overhead distribution systems. The proposed framework comprises three core modules to form a unified pipeline for autonomous field inspection: a 4D multi-modal map, Sparse-dense fusion network (SDFNet), and Rotational multi-pyramid Transformer with texture and augmentation (RoMP-Tax). The 4D multi-modal map aims to establish an accurate spatial-temporal representation of the maintenance area by integrating light detection and ranging (LiDAR), camera, inertial measurement unit (IMU), and global navigation satellite system (GNSS) measurements. The SDFNet detects telegraph poles from the 4D multi-modal map through geometry, pseudo, and fusion streams, which effectively extract both geometric and optical features. The RoMP-Tax, designed with a hybrid CNN–Transformer architecture enhanced by LBP-based texture encoding and Mixup augmentation, identifies insulators under complex textures and varying illumination. Extensive evaluations on field measurements and benchmark datasets demonstrate the high accuracy and consistent performance of the proposed framework with respect to multiple quantitative metrics, validating its robustness and generalizability. The proposed framework, deploying core technologies of the fourth industrial revolution, provides a reliable and efficient solution for estimating optimal bucket positioning, thereby contributing to the establishment of safe, data-driven live-line maintenance of distribution facilities.

KEYWORDS: Jumper line detection; live-line maintenance; sensor fusion; simultaneous localization and mapping; object detection

1 Introduction

Modern power distribution infrastructure constitutes a critical foundation for sustaining industrial productivity, social stability, and public safety in contemporary society. The uninterrupted supply of electrical energy supports the continuous operation of essential services, including transportation, communication, manufacturing, and residential utilities, thereby ensuring economic resilience and quality of life [1]. However, the structural components of distribution facilities, including conductors, insulators, and jumper lines, are continuously exposed to environmental loads such as mechanical stress, thermal expansion, and wind-induced vibration, which gradually lead to corrosion, fatigue, and joint loosening of power facilities [2].

These progressive degradations not only deteriorate the structural integrity of the distribution network but also increase the likelihood of the disruption of electric supply, caused by issues such as insulation failure, conductor separation, and service interruption [3–5]. Hence, periodic inspections and timely repairs are indispensable to sustaining the operational reliability of distribution networks and prevent unexpected power outages.

The conventional operation and maintenance (O&M) of overhead distribution facilities is primarily performed by electricians using aerial lift vehicles, which impose limitations on the safety, consistency, and operational efficiency of live-line maintenance. Specifically, operators repeatedly adjust the position and angle of an aerial bucket to secure an appropriate working posture and tool clearance because jumper lines are installed with irregular orientations around insulators and conductors. This bucket placement in close proximity to high-voltage power distribution facilities exposes workers to severe electrical hazards, fall risks, and mechanical interference, often resulting in accidents and occupational injuries [6,7]. Accurate bucket placement also depends heavily on the operator's experience, visibility, and situational awareness, suggesting that optimal bucket placement varies considerably under field conditions [8]. These difficulties suggest that automated maintenance of distributed networks could be effective; employing unmanned or sensor-assisted platforms may reduce manual intervention and enable remote operation in power transmission and distribution networks [9,10]. However, the application of unmanned systems to live-line maintenance should account for payload constraints and electrical hazards arising from their proximity to energized conductors [11,12]. While previous studies have remained a distinct challenge for active maintenance operations. This difficulty arises from the necessity of maintaining extreme spatial accuracy and robust sensor performance while navigating within the strong electromagnetic interference (EMI) fields generated by live distribution lines.

These limitations have motivated extensive research into multi-modal sensing methods for environmental cognition. Specifically, various multi-modal sensors, including optical cameras, light detection and ranging (LiDAR), inertial measurement units (IMUs), and global navigation satellite systems (GNSSs), have been developed for power infrastructure monitoring [13,14]. Each sensor supports complementary perceptual functions, including the geometric reconstruction of pole-mounted components and motion or positional estimation of the aerial bucket. Specifically, optical sensors provide detailed surface information that facilitates the assessment of contamination, cracks, and physical degradation [15]. LiDAR can offer accurate three-dimensional (3D) representations of surrounding structures [16]. IMU and GNSS measurements contribute to the precise estimation of motion and position, enabling the establishment of consistent spatial references across sequential viewpoints [17,18]. However, achieving a reliable perception of the environment in real time during live-line maintenance remains challenging. The reason is that the integration of multi-modal sensor measurements requires simultaneously ensuring consistent data acquisition and robust spatial awareness under diverse structural and environmental conditions.

Multi-modal mapping and localization frameworks have been extensively investigated by integrating LiDAR–IMU information to enhance spatial robustness [19,20]. These frameworks typically employ complementary sensor modalities to support real-time mapping, state estimation, and environmental cognition in complex outdoor environments. Specifically, LiDAR-based frameworks, including LiDAR odometry and mapping (LOAM) [21] and LiDAR inertial odometry via smoothing and mapping (LIO-SAM) [22] utilize the loosely coupled LiDAR–IMU integration method in which each sensor is processed independently before global registration. Although this configuration enhances computational efficiency, geometric–temporal misalignment remains unavoidable under dynamic motion or environmental interference [23,24]. Hence, temporal desynchronization among the LiDAR, camera, and IMU measurements results in the accumulation of spatial drift and distortion, particularly in environments with reflective conductors or irregular terrain.

To address this issue, new frameworks, including FAST-LIO [25], which has shown high efficiency in infrastructure mapping with rapid motion [25], and LINS [26], have been proposed. These methods jointly estimate system states by directly integrating LiDAR feature observations with inertial measurements through an iterated Kalman filter. Recent multi-modal frameworks have further extended this concept to LiDAR–camera–IMU fusion by jointly optimizing geometric and photometric constraints across modalities, including LiDAR-visual-inertial odometry via smoothing and mapping (LVI-SAM) [22] and Robust real-time RGB-colored LiDAR-inertial-visual tightly-coupled state estimation (R3LIVE) [27]. These methods significantly improve mapping precision and global consistency in complex environments. Nevertheless, accurate environmental reconstruction and cognition in distribution line environments require not only geometric consistency but also stable localization performance under complex visual and structural conditions. Hence, research on multi-modal localization should be further studied not only to adaptively synchronize LiDAR and optical observations but also to compensate for environment-induced drift in power-distribution environments.

Recognition of facility components also plays a critical role in autonomous maintenance. Recent studies have actively explored object recognition with optical sensors [28–31]. Specifically, neural networks (NNs) for object detection, including Faster region-based convolutional neural network [32], M2Det [33], and the YOLO family of detectors [34–37], have been proposed for detecting the poles and cross-arms of distribution networks due to their lightweight architectures and real-time inference capabilities. Despite effectiveness in general detection tasks, these conventional detectors exhibit significant challenges in the precise localization of inclined objects. These challenges are primarily attributed to the inherent high aspect ratio of inclined objects, which causes standard bounding boxes to incorporate excessive background noise. This noise results in a decrease in the Intersection-over-Union (IoU) accuracy and insufficient orientation awareness, thereby hindering the establishment of the precise spatial coordinates required for safe robotic operation [38]. To address these limitations, detection networks employing rotational bounding boxes, including Oriented R-CNN [39], RetinaNet [40], and Gliding Vertex [41], have been introduced, improving localization accuracy for inclined objects. However, these approaches still rely on CNN-only or Transformer-only architectures, restricting feature representation to either local geometric details or global contextual relationships [42,43]. These limitations suggest that further research should focus on vision-based NNs capable of jointly capturing local and distant features and robustly detecting inclined objects under geometrically complex and cluttered environments.

Two-dimensional (2D) object detection has limited accuracy and robustness because it relies solely on image information without considering depth information. This inherent limitation renders 2D object detection unsuitable for use in localization tasks for autonomous maintenance in large-scale distribution environments. Specifically, Dist-YOLO [44] and an extension of YOLOv3 [36] exhibit meter-level positional errors due to inherent depth ambiguity. These errors suggest that accurate distance-aware perception requires a new method for 3D object detection. Detection networks have been introduced to address this limitation with 3D point cloud data from LiDAR. Specifically, BtcDet was proposed to address occluded feature loss by addressing a Behind-the-Curtain method that reconstructs invisible regions for robust 3D detection [45]. PV-RCNN was also proposed by employing voxel-to-point feature fusion and keypoint refinement to enhance spatial localization accuracy [46]. A new neural network, titled PointNet, was proposed by directly learning geometric representations from unstructured point clouds without voxelization or projection [47]. These NNs would be more effective than conventional NNs for 2D object detection. However, the sparse characteristics of LiDAR still limit the capture of dense texture features that are essential for complex environments. Notably, overhead distribution facilities feature geometrically irregular and densely connected structures, suggesting that sparse geometric cues are insufficient for reliable detection [48]. Hence, object

recognition requires methods that can jointly extract both sparse geometric and dense texture features to ensure robust detection under real power distribution environments.

To overcome these limitations, this paper proposes a unified perception framework for the autonomous live-line maintenance of overhead distribution facilities. The proposed framework integrates 4D multi-modal map construction for global spatial registration, SDFNet-based pole detection for defining the region of interest, and RoMP-Tax-based insulator detection for isolated jumper line extraction. By leveraging information-level multi-modal fusion, the framework enables accurate, robust, and spatially consistent identification of jumper line positions, thereby providing essential spatial information for estimating the optimal bucket position during maintenance operations. The novelty and major contributions of this study are summarized as follows:

- The proposed framework integrates the inter-dependencies between 4D multi-modal mapping, SDFNet-based ROI definition, and RoMP-Tax-based insulator detection to fundamentally address the spatial inconsistency between global mapping and object detection. This integrated framework prevents the accumulation of independent errors from disconnected modules and provides a high-fidelity spatial foundation that moves beyond the limitations of standalone perception tasks.
- A new method is proposed to construct a unified 4D multi-modal map for live-line maintenance of overhead distribution facilities. This method executes an information-level mutual correction between IMU and GNSS information to enhance odometry and map stability. This method also employs an adaptive balancing method to handle the accuracy discrepancy between Visual-Inertial Odometry (VIO) and LiDAR-Inertial Odometry (LIO), thereby mitigating error propagation and ensuring consistent object localization within complex and feature-sparse structural environments.
- A multi-stream neural architecture, titled Sparse-dense fusion network (SDFNet), is introduced for utility pole detection from the 4D multi-modal map. SDFNet features a dedicated pseudo stream that aligns RGB features with LiDAR sparsity to address the intrinsic information imbalance in multi-modal fusion. These features are integrated by a neural network-based fusion stream, enabling robust pole localization even under sparse LiDAR information and varying illumination. This method secures reliability for establishing a precise region of interest (ROI) in complex field environments.
- A hybrid rotated-object detection neural network, titled Rotational multi-pyramid Transformer with augmentation and texture extraction (RoMP-Tax), is proposed to accurately detect insulators for extracting isolated jumper lines. The RoMP-Tax is specifically designed to leverage fine-grained texture features to detect objects from the highly complex and cluttered backgrounds of power-distribution environments. By incorporating a specialized texture-encoding module into the CNN–Transformer hybrid backbone, the neural network effectively utilizes additional object-specific characteristics that are often obscured in standard feature maps. This feature ensures the precise localization of high-aspect-ratio, inclined, and partially occluded components, which is critical for safe robotic intervention.

The remainder of this paper is organized as follows. Section 2 presents the proposed framework, including the overall procedure, construction of the 4D multi-modal map, and architectures of SDFNet and RoMP-Tax. Section 3 describes the experimental setup, including the hardware configuration of the aerial work platform, data-acquisition process, and neural network construction. Section 4 presents the details of the results and discussion, analyzing the accuracy of the 4D multi-modal map and the performance of SDFNet and RoMP-Tax. Finally, Section 5 concludes this paper by summarizing the major findings and discussing potential directions for future research on autonomous live-line maintenance of overhead distribution systems.

2 Proposed Framework

2.1 Entire Procedure

This subsection outlines the proposed framework for extracting jumper lines from live-line distribution facilities. The proposed framework comprises four phases for autonomous maintenance and repair with a robotic system. Prior to the four phases, the preprocessing phase synchronizes asynchronous inputs from various sensors, including LiDAR and GNSS at 10 Hz, a camera at 30 Hz, and an IMU at 120 Hz (Phase 0 in Fig. 1). The preprocessing phase utilizes the 120 Hz IMU timestamps as the primary temporal reference to align all other sensor data onto a unified clock. The framework projects the relatively sparse measurements from LiDAR, GNSS, and the camera onto this high-frequency reference grid through linear interpolation. The framework aligns measurements based on a single, high-frequency internal clock to compensate for the significant sampling rate discrepancies between sensors. The preprocessing phase estimates a measurement M_{target} a target time t stamp by interpolating between two adjacent measurements M_i and M_j recorded at t_i and t_j as follows:

$$M_{target} = M_i + \left(M_j - M_i\right) \frac{t - t_i}{t_j - t_i} \tag{1}$$

The implementation of the preprocessing phase ensures that multi-modal sensor measurements from LiDAR, camera, GNSS, and IMU remain strictly aligned regardless of the original sampling frequencies. This alignment is essential to confirm that the system maintains data consistency and spatial accuracy during the integration of multi-modal information. Following this pre-processing, the framework sequentially executes the four phases for jumper line detection.

First, a 4D multi-modal map is constructed by fusing LiDAR, camera, IMU, and GNSS measurements ((1) in Fig. 1). This phase aims to create an accurate spatial representation of the maintenance area. Specifically, the aerial lift car moves along its path while LiDAR, camera, IMU, and GNSS sensors spontaneously gather environmental information. Hence, the proposed system builds a global 4D map that includes telegraph poles and surrounding infrastructure. The detailed construction procedure of the 4D multi-modal map is described in Section 2.2.

Second, the real-time detection of the telegraph pole is executed in the distribution facilities ((2) in Fig. 1). This phase was examined to detect the telegraph pole using the SDFNet 3D object detection network. The detailed architecture and characteristics of the SDFNet are presented in Section 2.3. The locations of all telegraph poles are designated in the 4D multi-modal map and then the point-cloud data (PCD) of the telegraph pole of interest is extracted. This phase comprises two procedures: global frame transformation and coordinate aggregation. The former transforms the locations of telegraph poles detected in a single frame into coordinates defined in the global frame and then defines the geometric location of each pole with duplicate detections. This transformation is necessary because telegraph poles detected in a real-time process often result in duplicate detections owing to overlapping frames. Telegraph pole localization computes the global coordinates of each detected pole through a rigid-body transformation. The local detection results obtained using the SDFNet define each position in a single frame. The proposed method transforms these coordinates into the global frame using the estimated LiDAR-inertial pose, which consists of a rotation matrix R and translation vector T. The coordinate transformation is calculated as follows:

$$P_i^G = R \cdot P_i^L + T, \tag{2}$$

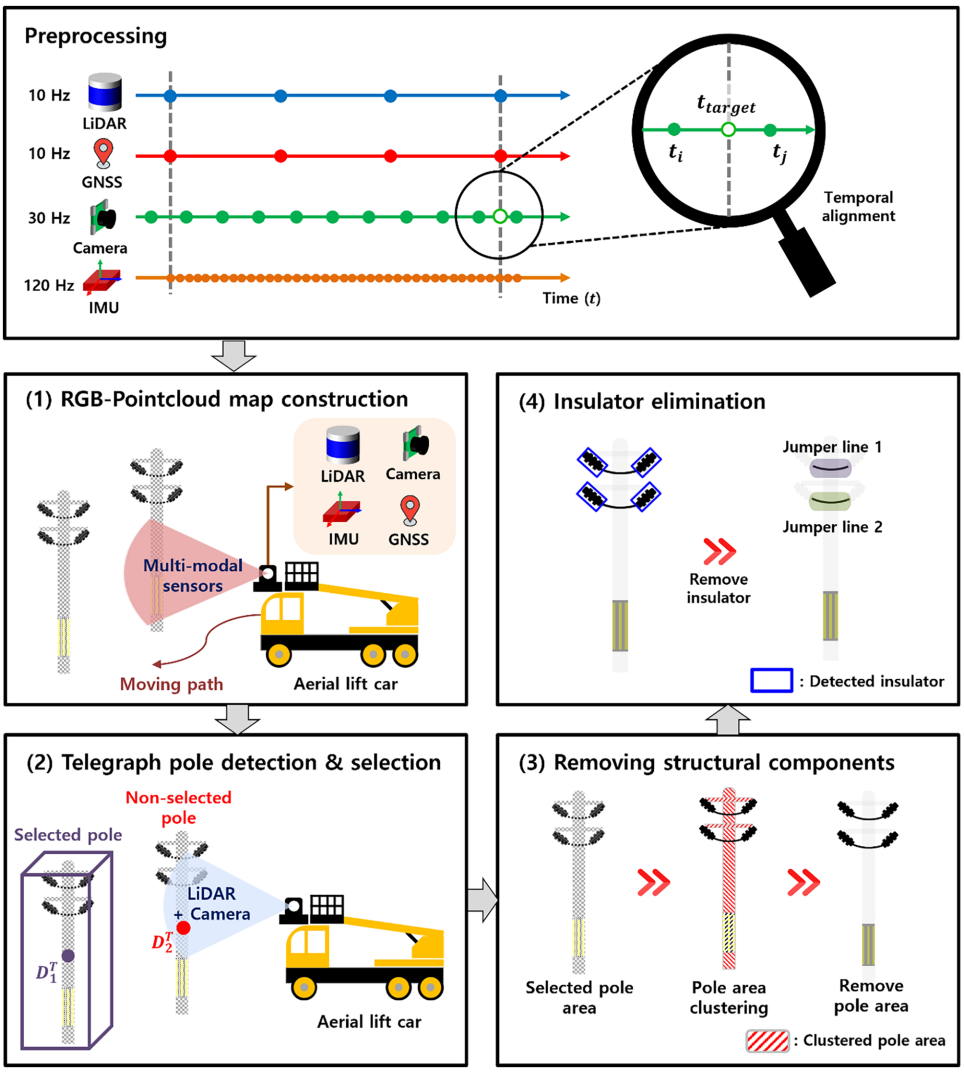

Figure 1: Schematic flowchart of the proposed framework for isolated jumper line detection.

where P_i^G and P_i^L denote the global position of the i_{th} telegraph pole and local coordinates detected by SDFNet, respectively. This transformation is essential because real-time detection frequently produces duplicate detections of the same pole owing to overlapping frames during continuous scanning. The coordinate aggregation procedure merges duplicate detections of the same telegraph pole into a single representative location within the global frame. To integrate duplicate detections of the same pole captured from varying viewpoints into a single global coordinate system, the proposed method utilizes an incremental k-d tree (iKD-Tree) [49]. This structure enables the real-time spatial indexing and dynamic updating of pole coordinates, which is crucial for maintaining spatial consistency as the aerial lift moves along the distribution line. The transformed point Tp_i retrieves nearby candidates within a predefined radius r. The candidate process is calculated as follows:

$$C_i = \{P \in \mathcal{T} \mid \|P - P_i^G\|_2 < r\}, \tag{3}$$

where C_i and $\mathcal{T}$ denote the set of candidate points located within radius r and the IKD-Tree storing previously registered pole locations, respectively. This process integrates the candidate telegraph pole coordinates into

a single centered geometric location to ensure consistent frames. The final integrated position is used to designate the corresponding telegraph pole region, which subsequently serves as the spatial reference for isolated jumper line extraction in later stages of the framework. Specifically, a subset of the point cloud is extracted within a three-meter range along the x and y axes centered at the unified telegraph pole location, to define the spatial region of interest. This localized point cloud region provides a bounded context for identifying jumper lines that are physically connected to the designated telegraph pole.

The third phase ((3) in Fig. 1) removes structural components, including both the vertical body and horizontal arms of a telegraph pole, through an unsupervised clustering algorithm. Specifically, DBSCAN identifies circular structures from the RGB point cloud projected onto the bird's eye view (BEV). DBSCAN classifies a point as a cluster member when the number of neighboring points satisfies the density condition with a predefined radius [50]. DBSCAN is formulated for telegraph pole clustering as follows:

$$N_\epsilon(p) \geq \{q \in D | \, dist(p,q) \leq \epsilon\}, \tag{4}$$

where $N_\epsilon(p)$, q and ϵ denote the set of neighboring points within a radius centered on point p, an individual point within the dataset D and the number of points required to form a dense area, respectively. The system filters out the pole body by extracting clusters that satisfy this criterion. The proposed method eliminates the remaining pole arms and attaches distribution lines using Mean shift clustering (MSC). These line structures exhibit elongated and linear shapes that differ geometrically from the circular pole body. MSC detects clusters by iteratively shifting candidate centers towards regions with higher local density. The MSC is formulated as

$$m(x) = \frac{\sum_{i=1}^{n} K(x_i - x) w_i x_i}{\sum_{i=1}^{n} KK(x_i - x) w_i}, \tag{5}$$

where $m(x)$, and $K(\cdot)$ denote the mean shift vector and the kernel function, respectively. The subscripts i denote the number of iterations and the number of PCD, respectively. This iterative process converges the cluster center to the local density peak, allowing the system to identify and eliminate elongated structures while preserving the jumper line.

The fourth phase ((4) in Fig. 1) eliminates the insulator in the distribution line to extract the isolated insulators using RoMP-Tax. Note that detecting insulators is difficult using only LiDAR because they appear small in the frame of the distribution line taken from an aerial lift car. Hence, this study incorporated texture information from optical images to detect insulators in the distribution line. The detailed architecture and characteristics of RoMP-Tax are described in Section 2.4. In addition, the isolated jumper line clustering groups each of the extracted isolated jumper lines into individual lines. This clustering process employs the MSC algorithm given by Eq. (4). The extracted isolated jumper line provides the structured spatial information necessary for decision-making. This isolated jumper line information enables continuous monitoring and allows the system to confirm the spatial positions of each jumper line in real time. The extracted jumper line positions assist in defining the optimal bucket location of the aerial work platform and automating bucket control operations.

2.2 Construction of 4D Multi-Modal Map

This subsection describes the detailed procedure for constructing the 4D multi-modal map ((1) in Fig. 1). This phase comprises six steps for integrating LIO and VIO.

Prior to describing the six steps, the proposed framework explicitly defines the coordinate systems and transformation notations to ensure spatial consistency and technical reproducibility. As illustrated in Fig. 2, five primary coordinate systems are employed: the global frame $\mathcal{F}_G$ defined in the east-north-up (ENU)

system, the body frame $\mathcal{F}_B$ attached to the inertial measurement unit (IMU) sensor, the LiDAR frame $\mathcal{F}_L$, the camera frame $\mathcal{F}_C$, and the pole frame $\mathcal{F}_P$ centered at the base of the target structure. The spatial mapping of a point from the frame $\mathcal{F}_j$ to frame $\mathcal{F}_i$ is represented by a 4 × 4 homogeneous transformation matrix T_{ij}, which is defined as:

$$T_{ij} = \begin{bmatrix} R_{ij} & t_{ij} \\ 0 & 1 \end{bmatrix}, \tag{6}$$

where R_{ij} and t_{ij} denote the 3 × 3 rotation matrix and 3 × 1 translation vector, respectively. In this framework, while the extrinsic parameters between the sensors, denoted as T_{BL} and T_{BC}, are pre-calibrated and remain static to ensure spatial consistency. Spatial alignment is achieved through a target-based optimization using a checkerboard, which transforms each sensor's coordinate system into the body frame. Furthermore, as described in Section 2.1, temporal registration is ensured by aligning sensor timestamps through linear interpolation. Based on this precise spatial and temporal calibration, the global trajectory $T_{GB}(t)$ is estimated online through the following procedure.

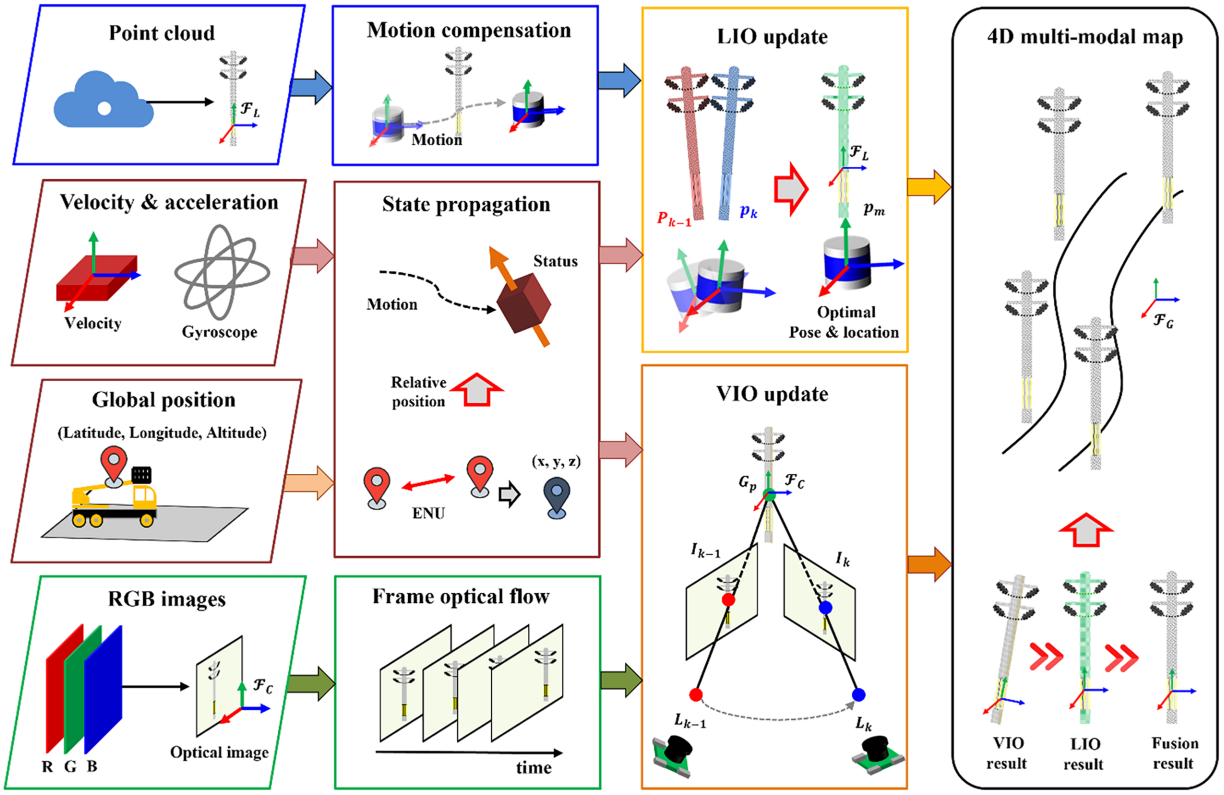

Figure 2: Flowchart of the 4D multi-modal map construction.

First, the motion of the mobile aerial work platform is estimated and compensated using LiDAR measurements (blue box in Fig. 2). Specifically, the compensated motion of the LiDAR point p_i^{corr} at timestamp ith is computed as follows:

$$\widehat{P_k} = T_k \cdot P_k, \tag{7}$$

where $\widehat{P_k}$, and T_k denote the raw point cloud data and the continuous-time sensor pose at timestamp k respectively. To ensure spatial consistency with the surrounding environment, the trajectory of the

continuous-time sensor must be refined using the iterative closest point (ICP) between the transformed LiDAR points at the source scan and the associated nearest neighbor in the target scan [51]. The relative motion constraint is formulated for two LiDAR scans at different timestamps as follows:

$$\triangle T = argmin \sum \left\| T \cdot P_{source} - P_{target} \right\| \tag{8}$$

where $\triangle T$ denotes the estimated relative transformation. The system incorporates ICP-based constraints into the trajectory to ensure geometric consistency between the source and target scans and then updates the LiDAR odometry with the newly registered PCD to refine the estimated sensor pose.

Second, the state of the mobile aerial work platform is propagated through IMU and GNSS measurements (brown box in Fig. 2). Specifically, the trajectory X is estimated by minimizing a cost function comprising the IMU and GNSS residuals as follows:

$$X = argmin \left(\sum_k \left\| r_{IMU,k} \right\|^2 + \sum_k \left\| r_{GNSS,k} \right\|^2 \right), \tag{9}$$

where $r_{IMU,k}$ and $r_{GNSS,k}$ denote the IMU and integrated residuals at timestamp k, respectively. The IMU residual $r_{IMU,k}$ is calculated from pre-integrated inertial measurements between timestamps k as follows:

$$r_{IMU,k} = \begin{bmatrix} R_k^T \left(p_{k+1} - p_k - v_k \Delta t - \frac{1}{2} g \Delta t^2 - \alpha_k \right) \\ R_k^T \left(v_{k+1} - v_k - g \Delta t \right) - \beta_k \\ \left(\log \left(R_k^T R_{k+1} \right) \right)^\vee - \theta_k \end{bmatrix}, \tag{10}$$

where R_k, p_k, v_k, Δt, g, α_k, β_k, and θ_k denote the rotation matrix at time step k, position at time step k, velocity, elapsed time between timestamps k and $k+1$, gravity vector in the global frame, pre-integrated position, velocity, and rotation increments computed from IMU measurements, respectively. Note that IMU-only integration accumulates bias over time, resulting in drift problems. GNSS provides ENU-referenced positions and is not affected by IMU integration drift [52]. Hence, GNSS and IMU measurements should be combined to compensate for the limitations of each sensor. The integrated residual $r_{GNSS,k}$ is defined as follows:

$$r_{GNSS,k} = p_k - z_{GNSS,k}, \tag{11}$$

where $z_{GNSS,k}$ denotes the relative position in the ENU frame from GNSS. Note that the combined minimization of the GNSS and IMU residuals secures a globally consistent trajectory because this approach effectively solves the long-term drift associated with IMU-only state propagation.

Third, the frame optical flow module (green box in Fig. 2) acquires consecutive RGB image frames to estimate the dynamic trajectory of the aerial work platform. Specifically, this step computes the temporal pixel displacement of visual features Δu_i, where the displacement of the i_{th} feature between timestamps t_k and t_{k+1} is computed as follows:

$$\Delta u_i = u_i(t_{k+1}) - u_i(t_k), \tag{12}$$

where $u_i(t_k)$ denotes the 2D pixel coordinates of the feature point at timestamp t_k. Hence, this process enables the proposed framework to track the visual features of optical images at each timestamp.

Fourth, the LIO is updated at each timestamp to construct the 3D geometric structure of the 4D multimodal map by tightly coupling LiDAR, IMU, and GNSS measurements (yellow box in Fig. 2). Specifically, the input LiDAR scan is registered by minimizing point-to-plane residuals, while motion distortion is compensated by addressing continuous-time trajectory reconstruction, which estimates backward-propagated

states by incorporating an error-state iterative Kalman filter (ESIKF) [25]. Specifically, the estimated state $\widehat{x_k}^{i+1}$ at iteration $i+1_{\text{th}}$ is computed as follows:

$$\widehat{x_k}^{i+1} = \widehat{x_k}^{i} \bigoplus \left(-Kz_k - (I - KH)\, P^{-1} \left(\widehat{x_k}^{i} \ominus \hat{x}_{k-1}\right)\right). \tag{13}$$

where K, H, P, $\widehat{x_k}^{i}$, and z_k denote the Kalman gain, Jacobian matrix, predicted state covariance, estimated state at the i_{th} iteration in frame k, and innovation residuals in the frame k, respectively. The subscripts i and k denote the iteration number and current timeframe, respectively. $\oplus$ and $\ominus$ denote the retraction and inverse retraction operators defined on the state manifold, respectively. This iterative process minimizes the overall estimation error by incorporating corrections from the sensor measurements into the state estimate, thereby enhancing the accuracy of the global 4D map.

Fifth, the VIO update (orange box in Fig. 2) enhances the visual map by rendering its texture and further refining the system state. This process minimizes two types of errors. The first error originates from the frame-to-frame point-to-plane (PnP) reprojection. The correction of this error aims to track accurate pose by projecting 3D map points onto the current image. Specifically, the frame-to-frame VIO minimization is as follows:

$$min \sum_{s=1}^{m} \left| r\left(x_k, \rho_{s,k}, G_p\right) + H_s \delta x_k \right|^2_{\sum s}, \tag{14}$$

where x_k, δx_k, $r(\cdot)$, H_s, $\rho_{s,k}$, and G_p denote the system state at frame k, error-state vector, residual function to compute the reprojection error between the observed 2D image points and projected 3D global map points, Jacobian of the partial derivatives of the residual function $r(\cdot)$ with respect to the error-state vector δx_k, measurement noise covariance of the projected 2D map points at frame k, and global 3D position of the points, respectively. The second error is the frame-to-map photometric error, which enforces texture consistency by minimizing cross-timestamp intensity differences and refining the predicted trajectory [27]. Photometric error minimization is formulated by projecting the transformed 3D global map points onto the current image frame and minimizing the intensity differences as follows:

$$min \sum_{i=1}^{N} \left| I_k\left(\pi\left(T_{kw} G_{p_i}\right)\right) - I_w\left(\pi\left(G_{p_i}\right)\right) \right|^2, \tag{15}$$

where N, I_k, T_{kw}, I_w, $\pi(\cdot)$, and G_{p_i} denote the number of selected points used for the photometric alignment, and image intensities at the current frame k, relative transformation matrix, reference frame w, projection function from the 3D space to 2D image plane, and 3D coordinates of the ith map point expressed in the global frame. Note that the error couples spatial projection and photometric residuals to enforce color consistency between reprojected map points and current-image pixels. Hence, minimizing the photometric error aligns the texture across frames and improves visual tracking and mapping accuracy.

Finally, the 4D multi-modal map is constructed by integrating the outputs of both the LIO and VIO updates (bold black box in Fig. 2). The frame optical flow estimates temporal pixel displacements and identifies matching feature points, yielding pixel observations u_i and u_j at timestamps t_i and t_j, respectively. The corresponding 3D point p_i^{cam} in the camera coordinate frame is calculated as follows:

$$p_i^{cam} = Triangular\left(u_i, u_j, T_{ci}^{cj}, K_{cam}\right), \tag{16}$$

where $Triangular(\cdot)$, T_{ci}^{cj}, and K_{cam} denote the process of back-projecting the pixel coordinates into normalized 3D direction vectors, relative transformation from the camera frame at time t_i, and camera

intrinsic matrix, respectively. The triangulated 3D point p_i^{cam} is then transformed into a global frame. The initial global position of point G_{pi}^{LIO} is estimated using the VIO-estimated camera pose as follows:

$$G_{pi}^{LIO} = T_{cw}^{LIO} \cdot p_i^{cam}, \tag{17}$$

where T_{cw}^{LIO} denotes LIO-refined camera pose. This initial camera estimate provides a coarse alignment of the map point within the global coordinate system to ensure subsequent refinement using the LIO update. The final 4D multi-modal map point $\widehat{G_{pi}}$ is calculated as follows:

$$\widehat{G_{pi}} = \begin{bmatrix} G_{p_i}^{LIO} \\ C_i \end{bmatrix} \in \mathbb{R}^6, \tag{18}$$

where G_{pi}^{LIO} and C_i denote the updated pose from LIO and the corresponding color, respectively. The RGB color corresponding to each 3D point is obtained by projecting the point onto the image plane, utilizing the intrinsic parameters of the camera and sampling the pixel value at the projected location as follows:

$$C_i = I_i \left(\pi \left(K_{cam} \cdot p_i^{cam} \right) \right), \tag{19}$$

where I_i denotes the projection function from 3D to 2D. This integration process combines the photometric information obtained from the VIO module with the geometric precision provided by the LIO update. Note that this 4D multi-modal map serves as the foundation for the subsequent phases of telegraph pole detection and jumper line extraction. Hence, the proposed framework not only achieves a high level of mapping precision but also ensures robustness to challenging environmental conditions.

2.3 SDFNet

This subsection describes the SDFNet, which detects the telegraph pole by analyzing the LiDAR and optical features of the second phase ((2) in Fig. 1). The proposed neural network features three key characteristics: geometry, pseudo, and fusion streams.

First, the geometry stream (① in Fig. 3) extracts geometric features from 3D PCD, which are structured as $\mathbb{R}^{N\times 3}$, where N denotes a number of PCD. The inherent sparseness of PCD necessitates feature extraction to be designed such that geometric consistency is retained while spatial resolution is enhanced. This study addresses the architecture of multi-scale feature extraction (MS extraction) to overcome this spatial resolution limitation. MS extraction employs multi-scale layers to extract various geometric features under dynamic environmental conditions. Specifically, geometric feature maps F_1^g, F_2^g, F_3^g, and F_4^g are extracted as follows:

$$F_k^g = \mathcal{F}_{MS} \left(P, \theta_k \right), k \in \{1, 2, 3, 4\}, \tag{20}$$

where $\mathcal{F}_{MS}(\cdot)$, P, and θ_k denote the multi-scale feature extraction, input PCD, and learnable parameters in the geometry stream, respectively. The subscript k denotes the position of the kth layer within the network. The number of MS was determined as four through Bayesian optimization (BO) to ensure accuracy and robustness. Then, multi-scale fusion (MS fusion of ① in Fig. 3) is executed to integrate each feature map to the unified resolution and concatenate along the channel dimension as follows:

$$F_g^{MS} = Concat \left(F_{g,1}, F_{g,2}, F_{g,3}, F_{g,4} \right), \tag{21}$$

where F_g^{MS} and $Concat(\cdot)$ denote the resulting multi-scale geometric feature map aggregated from various geometric feature maps and concatenation along the channel dimension, respectively. The integrated

multi-scale geometric feature map retains distinct features that preserve both global and local geometric characteristics.

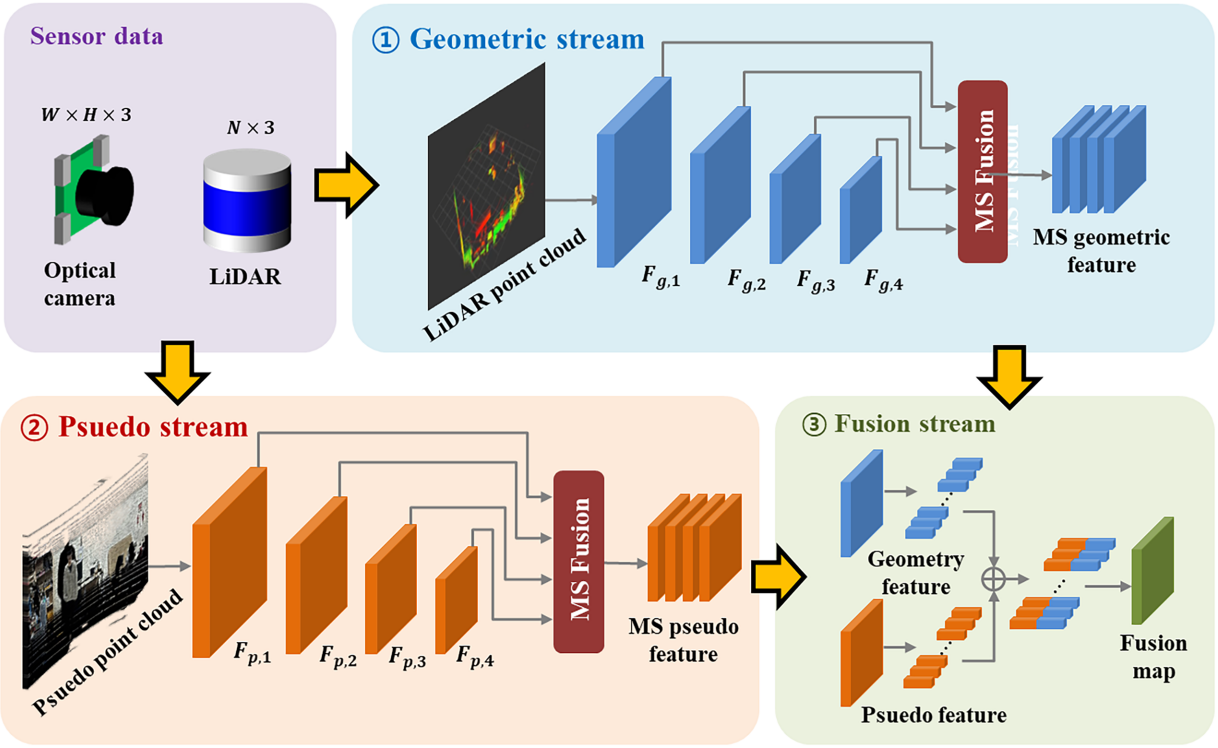

Figure 3: Architecture of SDFNet for telegraph pole detection.

Second, the pseudo stream (② in Fig. 3) extracts features from pseudo PCD by combining PCD and RGB images. The RGB images are structured as $\mathbb{R}^{w\times h\times 3}$, where superscript w and h denote the pixel width and height, respectively, along with depth information. Note that the RGB texture in pseudo PCD enables the network to incorporate dense photometric and spatial information that is not available in geometric features [53]. Each geometric map g_i encodes the real-world 3D coordinates at each spatial location in the image domain as follows:

$$g_i\left(u,v\right)=\left[x,y,z\right]^T\in\mathbb{R}^3, \tag{22}$$

where $g_i(u,v)$ and $\left[x,y,z\right]^T$ denote the 3D spatial position corresponding to the image coordinate and the lateral, vertical, and depth components of the geometric point, respectively. The associated RGB values are extracted from image I_i and concatenated with the spatial position to construct a colored pseudo point $\widetilde{p}_i$ as follows:

$$\widetilde{p}_i=Concat\left(g_i\left(u,v\right),I_i\left(u,v\right)\right)\in\mathbb{R}^6, \tag{23}$$

where $I_i(u,v)$ denotes the RGB color vector obtained from an image. The set of all pseudo point clouds P_{psuedo} is represented as follows:

$$P_{psuedo}=\{\widetilde{p}_i\}_{i=1}^{N}, \tag{24}$$

where N denotes the total number of pseudo points generated from a single frame. The constructed pseudo PCD are passed through a multi-scale feature extractor to obtain hierarchical features at different scales of $F_{p,1}$, $F_{p,2}$, $F_{p,3}$, and $F_{p,4}$ as follows:

$$F_{p,k} = \mathcal{F}_{MS}\left(P_{psuedo}, \phi_k\right), k \in \{1, 2, 3, 4\}, \tag{25}$$

where P_{psuedo} and ϕ_k represent the dense RGB features and learnable parameters in the pseudo stream, respectively. The number of MS layers in the pseudo stream is also optimized through BO to ensure the accuracy and robustness of 3D object detection. To enable feature fusion across different spatial resolutions, each feature map is resized to a unified resolution and concatenated along the channel dimension as follows:

$$F_p^{MS} = Concat\left(F_{p,1}, F_{p,2}, F_{p,3}, F_{p,4}\right), \tag{26}$$

where F_p^{MS} denotes the resulting pseudo feature map that aggregates hierarchical geometric and photometric features. The aggregated pseudo-features are also refined through MS extraction, thereby leveraging both geometric and texture information and enhancing the spatial features.

Finally, the fusion stream (③ in Fig. 3) fuses geometric and pseudo features to construct a comprehensive multimodal map that incorporates both sparse LiDAR PCD and the dense RGB image. This fusion process specifically addresses the sparse-dense imbalance inherent in LiDAR-camera systems by compensating for the low resolution of LiDAR PCD with dense RGB texture. In the context of telegraph pole detection, this architectural design is crucial because the proposed architecture of SDFNet preserves the fine structural boundaries of slender objects that sparse LiDAR points alone may fail to capture. Specifically, global average pooling (GAP) is applied to both streams to reduce the spatial dimensionality while retaining a global summary of each activation map. The vectors processed by the GAP $g_g \in \mathbb{R}^{C_g}$, $g_p \in \mathbb{R}^{C_p}$ are calculated as follows:

$$\begin{aligned} g_g &= \frac{1}{hw}\sum_{i=1}^{h}\sum_{j=1}^{w}\left(F_g^{MS}\right)_{i,j} \in C_g, \\ g_p &= \frac{1}{hw}\sum_{i=1}^{h}\sum_{j=1}^{w}\left(F_p^{MS}\right)_{i,j} \in C_p, \end{aligned} \tag{27}$$

where h and w denote the spatial height and width of each feature map, respectively, and C_g and C_p denote the numbers of channels in the multi-scale geometric and pseudo feature maps, respectively. The final fused feature F_{fusion} is obtained through weighted feature aggregation as follows:

$$F_{fusion} = \sigma\left(W_{FC} \cdot Concat\left(g_g, g_p\right) + b_{FC}\right), \tag{28}$$

where σ, W_{FC}, and b_{FC} denote the activation function, weight matrix, and bias, respectively. Note that SDFNet effectively integrates multi-modal features by leveraging spatial and structural characteristics from both sources. Specifically, geometric features contribute to precise spatial positioning, whereas the pseudo features incorporate dense texture information derived from RGB images.

2.4 RoMP-Tax

This subsection describes the detailed architecture and key characteristics of RoMP-Tax, which is effective for the insulator elimination phase ((4) in Fig. 1). Fig. 4 illustrates the architecture of the proposed RoMP-Tax. Note that the RoMP-Tax has four key characteristics: a rotational bounding box, an augmentation-texture feature learning strategy, a multi-scale multi-level (MSML) feature extraction module, and a pyramid vision transformer (PVT). These four characteristics are described in detail below.

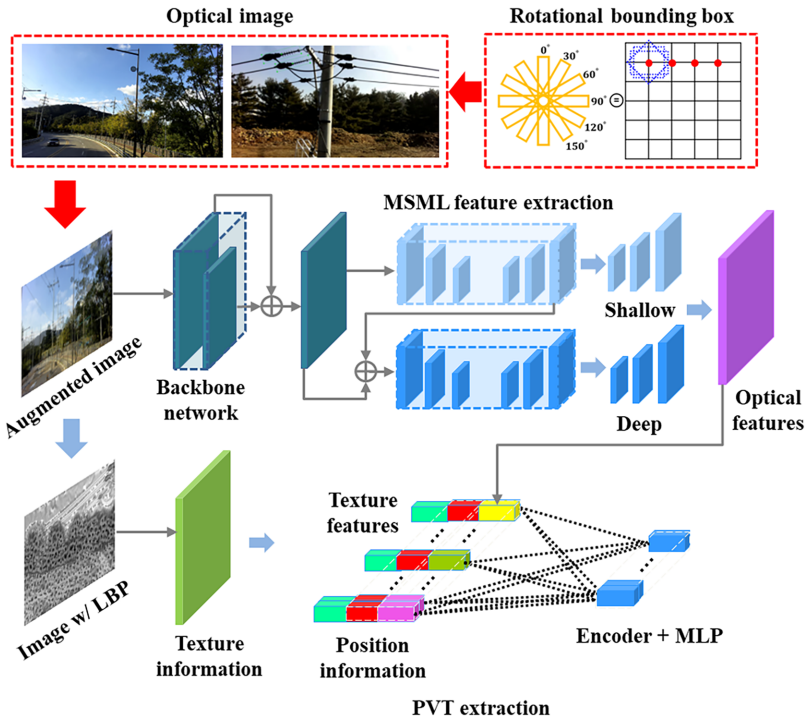

Figure 4: Architecture of RoMP-Tax for insulator detection.

First, RoMP-Tax detects the object of interest through a rotational bounding box. The rotational bounding box designates the inference region of interest with five parameters: coordinates of a center point x_c and y_c, width ω, height h, and angle α. It is worth noting that RoMP-Tax incorporates an angle parameter α to effectively detect the object of interest regardless of the inclination angle with respect to the background and the high aspect ratio of objects. Specifically, a rotational bounding box has significant advantages in accurate labeling and object detection because the angle α accounts for the sophisticated object area regardless of the inclination angle of the object of interest with respect to the background. This feature also minimizes features extracted from the background when executing object detection because the rotational bounding box constrains the inference region in accordance with the geometric orientation and structural boundaries of the target, improving the robustness for test images [38,54]. In comparison with segmentation-based methods, constructing the ground truth by addressing rotational bounding boxes requires substantially less annotation effort while maintaining comparable detection accuracy. In other words, the rotational bounding box secures similar accuracy to segmentation, with minimal effort to prepare a large amount of labeled data for supervised learning [55,56], suggesting that this feature is effective in terms of the accuracy, cost, and effort of labeling.

Second, the augmentation-texture feature learning strategy incorporates Mixup data augmentation and local binary pattern (LBP) for texture analysis [57,58]. This feature aims to further enhance the insulator detection accuracy compared to RoMP-T. Specifically, Mixup augmentation generates augmented training samples by linearly interpolating both the pixel values and corresponding labels of two randomly selected images. The augmented image $\overline{sp}$ obtained by combining sample pairs sp_i and sp_j is calculated as

$$\overline{sp} = \lambda \cdot sp_i + (1 - \lambda) \cdot sp_j, \tag{29}$$

where λ denotes the interpolation coefficient. The interpolation coefficient λ is determined through hyperparameter optimization, as described in Section 3.3. Hence, the Mixup augmentation improves the generalizability of the detector by constructing diverse training data and reducing the risk of overfitting. In addition, LBP aims to account for fine-grained textural details by encoding local intensity variations through a comparison between each pixel and its surrounding neighbors, thereby improving feature representation. The LBP between a central pixel $pixel_C$ and the surrounding neighbors $pixel_P$ is calculated as follows:

$$LBP_{P,R}\left(x_c, y_c\right) = \sum_{p=0}^{P-1} s\left(pixel_P - pixel_c\right) \cdot 2^P, \tag{30}$$

where $LBP_{P,R}(\cdot)$, x_c, y_c, $s(\cdot)$, $pixel_P$, and $pixel_c$ denote the LBP operator of the central pixel, x coordinate of the central pixel, y coordinate of the central pixel, thresholding function, intensity values of the neighboring pixels, and intensity value of the central pixel, respectively. The subscripts P and R denote the total number of neighboring pixels and radius of the neighborhood, respectively. This process encodes local structural patterns into a binary representation to improve the discriminative capability of the extracted features.

Third, the MSML feature extraction module employs a multi-level convolutional autoencoder architecture. The detailed formula for extracting the MSML features is as follows:

$$F_{l,k} = \mathcal{F}_l\left(\mathcal{F}_{MS}\left(X_{base}, x_k^{l-1}\right)\right), l \in \{1,2\}, k \in \{1,2,3,4\}, \tag{31}$$

where $F_{l,k}$, $\mathcal{F}_l$, X_{base}, and x_k^{l-1} denote the extracted MSML features, the multi-level feature extractor, feature with the i_{th} scale in the $l-1_{\text{th}}$ convolutional autoencoder, base feature, and l_{th} convolutional operation processing, respectively. Subscripts l and k denote the level and scale of the convolutional autoencoder, respectively. These extracted MSML feature maps contain semantic and distinct features of different sizes and levels of complexity. The MSML feature fusion module then combines several features extracted through concatenation and elementwise computation-based 1×1 convolution $(1 \times 1conv)$. This concatenated feature map F_{concat} is calculated as follows:

$$F_{concat} = Concat\left(\mathcal{F}_{(l,k)}\right), l \in \{1,2\}, k \in \{1,2,3,4\}, \tag{32}$$

where $Concat(\cdot)$ denotes the concatenation operation. Concatenated features are then condensed through an elementwise operation on the channel axis by executing $1 \times 1conv$, resulting in the final MSML feature map. This characteristic ensures high prediction accuracy and robustness when detecting the insulator because incorporating several features at different sizes and levels is effective in constraining the feature map in consideration of the operational conditions of jumper line detection.

Fourth, RoMP-Tax utilizes the PVT to extract the global correlation between all pixels. The PVT includes three steps: patch embedding, position embedding, and Transformer encoder. Patch embedding refers to grouping a set of pixels into one unit and treating them as one pixel in a 2D image. Specifically, the patch embedding in the PVT reshapes the image $x \in \mathbb{R}^{h \times w \times c}$ into $\mathbb{R}^{n \times (Patch^2 c)}$, where H, W, c, $Patch$, and n denote the height and width of the input image resolution, number of channels, size of patch, and total number of patches, respectively. The positional embedding allows operation with 2D position information in MLP. Specifically, the positional embedding in the PVT is calculated as

$$\begin{aligned} PE_{(2i)} &= sin\left(pos/10000^{\frac{2i}{d_{model}}}\right), \\ PE_{(2i+1)} &= cos\left(pos/10000^{\frac{2i}{d_{model}}}\right), \end{aligned} \tag{33}$$

where PE, pos, and d_{model} denote the value of positional encoding, position of the patch, and dimension of the PVT model in RoMP-Tax, respectively. The subscript i denotes the overall dimension of the flattened features. This process prevents the loss of location information in 2D through a flattening process in the PVT. The Transformer encoder performs attention operations by executing key k, query q, and value v, denoting the main pixel value, the set of pixels providing information, and the semantic result for the key, respectively. The three values of the key k, query q, and value v are calculated using parameters of W_k, W_q, and W_v, where W_k, W_q, and W_v denote the query, key, and value matrices, respectively. Finally, a new feature map is constructed in the Transformer encoder through the interaction of key k, query q, and value v through attention. Specifically, attention in the PVT with spatial reduction attention (SRA) is calculated as

$$Attention\,(q,k,v) = softmax\left(\frac{qk^T}{\sqrt{d_{head}}}v\right), \tag{34}$$

where $Attention(\cdot)$, $softmax(\cdot)$, and d_{head} denote the SRA process, softmax function, and dimension of head in RoMP-Tax, respectively. This SRA method significantly reduces the computational cost of PVT, similar to other vision-Transformer-based neural networks (Wang et al., 2021), thereby optimizing computational efficiency.

3 Experiments

This section describes the experimental setup designed to validate the proposed method for jumper line extraction. This section comprises three subsections, describing hardware configuration, dataset acquisition, and neural network construction. First, the hardware system integrating high-precision sensors with a dedicated processing unit is presented in Section 3.1. This hardware system is designed not only for real-time cognition of environments but also for the computation of accurate feature extraction to facilitate robust detection of the overhead distribution facilities of interest in complex environments. Second, the dataset acquisition is described in Section 3.2. This subsection describes the collection of field data from overhead distribution facilities. This subsection also presents the KITTI-3D benchmark dataset [59] and dataset for object detection in aerial images (DOTA-v2.0) [60], which is used for performance evaluation of the proposed framework. By combining practical field measurements with a widely used benchmark, the evaluation considers both the accuracy and robustness of 3D object detection. Third, the neural network construction is presented in Section 3.3. This subsection describes the setup of the learning framework, where SDFNet and RoMP-Tax are implemented to refine the feature representation and improve detection accuracy. Note that the architecture of the proposed neural network aims to ensure stable operation across diverse environments, where structural variability and visual complexity may degrade performance. The incorporation of these neural networks establishes the algorithmic core of the proposed framework for jumper line extraction.

3.1 Hardware Configuration of the Aerial Work Platform

This subsection describes the hardware configuration designed for data acquisition and real-time processing (Fig. 5). The system consists of five primary components: LiDAR, RGB camera, IMU, GNSS module and a signal processing unit. Table 1 summarizes the technical specifications of these components. LiDAR measures the 3D geometry of the environment to satisfy the tolerance required for distribution facility inspection. The RGB camera captures visual appearances to compensate for sparsity in LiDAR measurements. IMU enables continuous state propagation for stable localization when global updates are unavailable. The GNSS module provides absolute positions to minimize cumulative drift through real-time kinematic correction. The signal processing unit executes high-throughput tasks to ensure real-time

cognition and map construction. The hardware configuration is determined by minimum performance criteria to ensure framework adaptability for live-line maintenance.

Table 1: Hardware specifications for the aerial work platform.

Components	Model	Specifications
LiDAR	Velodyne Alpha Prime [61]	128 channels, 245 m range, ±3 cm accuracy
Optical camera	e-con Systems See3Cam [62]	1920 × 1080 @ 30 fps, 120° DFOV
IMU	VectorNav VN100 [63]	0.5° (Roll/Pitch), 2.0° (Yaw) accuracy
GNSS	Ublox ZED-F9P [64]	2 cm accuracy (RTK support)
Processing unit	Daven V200	i7-13700K CPU, 64 GB RAM, RTX 3070 GPU

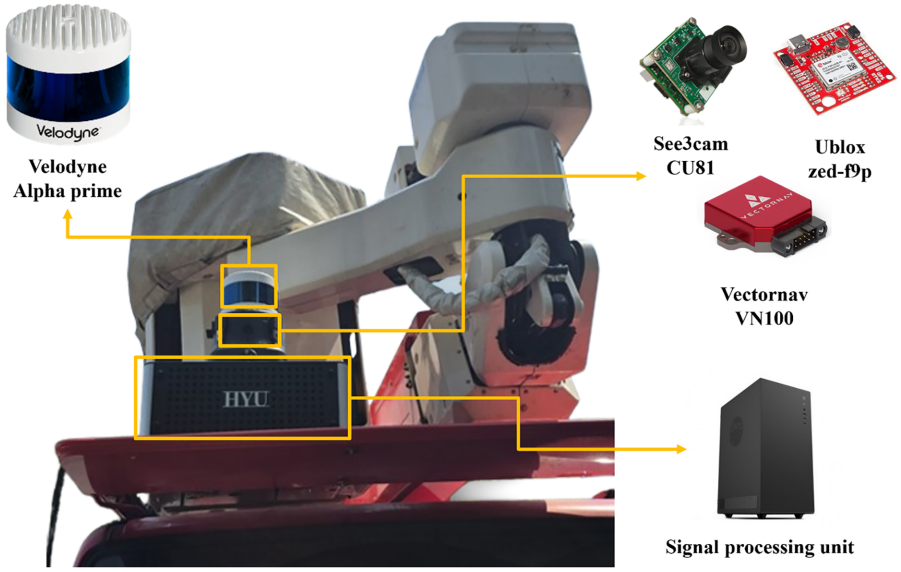

Figure 5: Hardware configuration of the aerial work platform, comprising LiDAR, optical camera, IMU, GNSS, and signal processing unit.

3.2 Data Acquisition

This study utilizes a field dataset and public benchmarks to evaluate the proposed framework [59,60]. Field measurements from distribution line experiments verify the jumper line detection procedure under real-world conditions. To ensure a robust evaluation of these multiple tasks, we constructed a field dataset comprising five scenarios with a total duration of 15 min and 26 s (Table 2). The KITTI-3D and DOTA-v2.0 benchmark datasets evaluate the 3D and 2D object detection performance of the SDFNet and RoMP-Tax.

First, the field experiments were conducted on three sites of distribution lines in South Korea (Fig. 6): Gongneung-dong in Nowon-gu (GN), Munji-dong in Yuseong-gu (MY), and Songgang-dong in Yuseong-gu (SY). Specifically, the GN site was located on flat terrain with minimal obstacles between utility poles, providing a controlled environment for baseline evaluation (Fig. 6A). The MY site, located on uneven ground, served as an intermediate case for validating the robustness of the proposed framework under irregular terrain conditions (Fig. 6B). The SY site is an energized distribution line located in a mountainous region (Fig. 6C), representing a real live-line maintenance environment. Hence, these different characteristics of the three sites are effective for evaluating the accuracy and robustness of the proposed framework in diverse

environments. The five datasets in Table 2 were strategically selected to encompass two operational statuses, two line types, and three different terrain characteristics. Specifically, the entire dataset was acquired to include jumper lines, insulators, and telegraph poles in overhead distribution facilities. Datasets #1 to 4 were measured on de-energized lines disconnected from the electrical grid to ensure operational safety of the experiments, whereas dataset #5 was measured on energized lines to evaluate the proposed framework under actual operational conditions, i.e., live-line conditions. The datasets from the field experiments included a brief overview of the experiments, including the status of the aerial lift car, site locations, current state of the distribution line, and experimental dates. These datasets also included 3D PCD from a 128-channel LiDAR sensor, RGB images from an RGB camera, inertial measurements from an onboard IMU, and absolute positioning from a GNSS receiver. All measurements were time-synchronized, suggesting that the experiments recorded multimodal measurements at the distribution facilities. The proposed framework was evaluated based on recorded measurements, and the detailed results are presented in Section 4.

Table 2: Experimental setup and conditions of field sites.

No.	Aerial Lift Car Status	Site Name	Line Type	Terrain Characteristic	Date	Duration
#1	Static	GN	De-energized line	Flat terrain	24.07.31	1 m 15 s
#2					24.08.14	1 m 25 s
#3					24.09.04	1 m 28 s
#4	Driving	MY	De-energized line	Irregular terrain	24.01.22	3 m 37 s
#5	Driving	SY	Energized line	Mountainous	24.09.25	7 m 41 s
Total	2 Statuses	3 Sites	2 Types	3 Terrains	5 Days	15 m 26 s

Figure 6: Representative images of the three experimental sites: (**A**) GN, (**B**) MY, and (**C**) SY.

Second, two public benchmark datasets are further utilized to evaluate the object detection performance [59,60]. The KITTI-3D dataset [59] provides a standard benchmark to compare SDFNet through optical images and LiDAR point clouds [46,47]. The DOTA-v2.0 dataset [60] also validates the 2D object detection performance of RoMP-Tax. This dataset contains aerial images with rotational bounding box annotations to evaluate rotation-aware detection under diverse scenarios. These benchmarks ensure a robust validation of the proposed detection capabilities.

3.3 Neural Networks Construction

This subsection describes the construction of SDFNet and RoMP-Tax, which were designed to address the challenges of jumper line detection in overhead distribution facilities. Four NVIDIA Tesla A100 GPUs and two AMD EPYC™ 7543 CPUs were employed for training, validation, and testing of two neural networks. The utilization of data parallelism across GPUs accelerated the training process while ensuring stable convergence.

SDFNet integrates convolutional and fully connected layers for local and global representation learning because it efficiently processes both point cloud and optical data in real time. Specifically, the architecture of SDFNet comprises three main components: a geometric stream, a pseudo stream, and a fusion stream. The geometric stream consisted of four 3D convolutional layers with $3 \times 3 \times 3$ kernels, and the number of channels was progressively increased from 32 to 256. Each convolutional layer was scaled by a factor of 1/2 for hierarchical feature extraction. The pseudo stream was constructed in a similar manner by employing four 3D convolutional layers with $3 \times 3 \times 3$ kernel to process RGB-depth data. This architecture ensures high feature fidelity during multi-modal fusion, thereby strengthening the robustness of the neural network in complex inspection environments.

RoMP-Tax integrates convolutional and Transformer-based layers to extract both local and correlation features effectively. Specifically, this architecture consists of two autoencoders, each comprising four encoder and decoder layers. Each feature map was scaled by a factor of 1/2 with respect to the layers of the CNN in the encoder, whereas each feature map was scaled by a factor of two with respect to the layers of the CNN in the decoder. Hence, each autoencoder composed of a CNN effectively extracted multi-scale features. A series connection of the two autoencoders also yielded multi-level feature maps, implying that MSML feature maps were generated by multi-level autoencoders. In the transformer, the patch size of the first stage was set to five to reduce computational burden. Note that the first stage required significant memory capacity, which was alleviated by applying a larger patch size. Each feature map was scaled by a factor of 1/5 with respect to the first layer and by a factor of 1/2 with respect to the other layers of the Vision Transformer.

The authors directly collected the datasets for training, validation, and testing through field experiments at the GN, MY, and SY sites. This self-captured data comprises annotated telegraph poles, insulators, and vegetation. Table 3 summarizes the dataset configurations for both SDFNet and RoMP-Tax to provide details on input types and instance numbers. SDFNet utilizes multi-modal data of LiDAR-based PCD and optical images to train and evaluate the 3D object detection performance. RoMP-Tax employs optical images to train and evaluate the 2D object detection performance. The detailed descriptions and experimental setups regarding the hyperparameter optimization for these models are provided in the Supplementary Materials.

Table 3: Dataset configurations for SDFNet and RoMP-Tax.

Neural Network	Data Type	Instances	Data		
			Train	Validation	Test
SDFNet	3D PCD & Optical image	Telegraph pole	1546	236	325
		Vegetation	919	108	176
RoMP-Tax	Optical image	Insulator	3717	531	1062

4 Results and Discussion

This section presents the results of the field experiments and benchmark datasets with an in-depth discussion. First, the detailed results from each phase are demonstrated using field measurements to verify the feasibility of the proposed framework. Second, the accuracy of the 4D multi-modal map, which was constructed by integrating LiDAR, camera, GNSS, and IMU data, was validated under field conditions to ensure precise coordinate registration and geometric alignment. Third, the SDFNet detection of telegraph poles was evaluated using PCD and optical images from the field experiments. The robustness of SDFNet was further validated on the KITTI-3D dataset under urban driving environments with diverse viewpoints [59]. Fourth, the RoMP-Tax insulator detection was evaluated using optical images only from field experiments. The robustness of RoMP-Tax was also validated on the DOTA-v2.0 dataset using satellite images with rotated instances and complex backgrounds [60]. Because direct comparison of the entire proposed framework with existing approaches is not feasible, the individual modules were validated on widely used benchmark datasets, which enabled objective comparison of both neural networks with previous approaches. Most previous studies have focused on single-modal sensing or partial anomaly detection.

4.1 Entire Procedure

This subsection describes the detailed results of the proposed framework comprising four sequential phases (Fig. 7): 4D multi-modal map construction, telegraph pole detection, structural component removal, and insulator elimination.

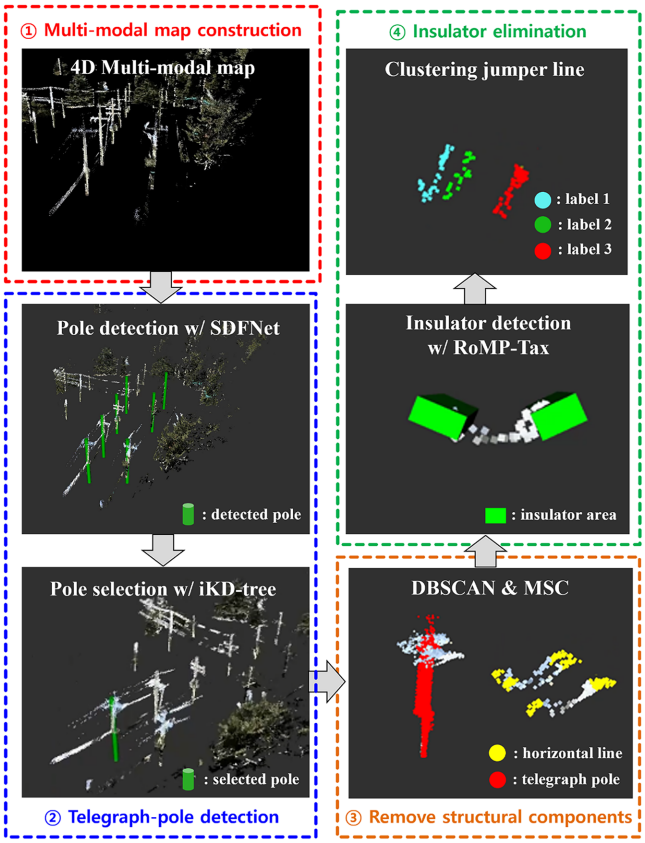

Figure 7: Results from the proposed framework for jumper line detection.

In the first phase (① in Fig. 7), the proposed framework successfully generated a 4D multi-modal map by combining LIO and VIO with GNSS measurements and optical images. Specifically, the multi-modal sensor data were time-synchronized and consistently aligned into a unified global coordinate frame, enabling precise representation of telegraph poles and surrounding facilities. Note that LIO ensured stable trajectory estimation during long-range motion but had the limitation of reduced accuracy in local trajectory estimation. In contrast, VIO effectively provided precise short-term refinements through image-based updates but had the limitation of drift accumulation over extended trajectories. By maximizing the advantages of LIO and VIO, while compensating for the disadvantages of LIO and VIO, the proposed framework significantly improved the overall mapping accuracy. The integration of geometric and visual information, as illustrated in ① in Fig. 7, provides a comprehensive representation of the distribution facilities, including their surrounding environment. This outcome established a reliable foundation for subsequent phases.

In the second phase (② in Fig. 7), telegraph poles were detected in real-time using SDFNet, which utilizes voxelized PCD and optical images. Specifically, SDFNet employed parallel convolutional streams to process the two modalities and fully connected layers to extract the fused features to identify pole structures. The detected poles were then transformed into the global coordinate frame using the estimated LiDAR inertial pose $T(t_i)$, ensuring consistency across overlapping frames. Duplicate detections caused by continuous scanning were merged into unified pole positions by applying an iKD-Tree. As illustrated in ② in Fig. 7, the green-colored cylinders visualize the registered pole coordinates, verifying that the estimated positions almost coincide with the actual telegraph poles. In the pole selection, a single telegraph pole was uniquely retained in the global frame from the electrician, thereby ensuring stable and unambiguous pole registration. The extracted PCD of the selected pole area enabled jumper line detection by isolating the telegraph pole from the surrounding environment for subsequent phases.

In the third phase (③ in Fig. 7), structural components of the telegraph pole were eliminated from the extracted pole area detected by SDFNet to isolate candidate line structures. Specifically, the horizontal and vertical components of the poles were removed by applying the DBSCAN unsupervised clustering method. This method was applied to the projected PCD, where MSC was used to extract elongated linear structures that correspond to distribution lines. This removal process excluded the main pole structures while preserving only the candidate jumper line and insulator segments. Fig. 7 ③ clearly depicts the successful removal of the horizontal and vertical components, confirming that this phase provided a refined structural representation in which only the target line components were retained for the subsequent insulator elimination phase.

In the fourth phase (④ in Fig. 7), insulators were first detected by RoMP-Tax, which incorporates both convolutional and Transformer-based layers to enhance spatial and contextual reasoning. Optical images were incorporated because the insulators represented in the PCD in the LiDAR-only frames have a limited size and resolution. RoMP-Tax enabled the reliable identification of insulator regions with optical images, and MSC effectively clustered the isolated jumper lines into individual groups. Hence, the insulator areas were clearly detected and removed, as shown in Fig. 7 ④, verifying that this phase effectively provided structured spatial information on jumper line locations. Note that MSC enabled the candidate jumper lines to be clearly clustered into individual groups, thereby facilitating the assignment of precise work locations in subsequent maintenance tasks.

The experimental results include an ablation study to clarify that each sensor modality contributes to the overall system performance (Table 4). Specifically, the performance of the proposed 4D multi-modal map was compared with two baseline configurations: a LiDAR-only map (w/o Camera), utilizing only LIO updates, and a Camera-only map (w/o LiDAR), utilizing only VIO updates. Remarkably, the comparison demonstrates that the proposed 4D multi-modal framework achieved the highest accuracies of 94.4%,

100.0%, and 93.8% at the sites of GN, MY, and SY, respectively, yielding an overall average accuracy of 96.1%. These results confirm that the multi-modal approach outperforms the LiDAR-only configuration by an average of 7.3% and the camera-only configuration by a substantial margin of 37.1%. These results also indicate that the 4D multi-modal map significantly outperforms the single-modality baselines in that the integration of LiDAR and camera data compensates for the limitations inherent in individual sensors. The missed detections in the 4D multi-modal map were confined to regions not captured during sensor data acquisition because occlusions from overlapping structural components and oblique sensor perspectives hindered the camera and LiDAR from acquiring measurements of certain jumper lines. These occlusions primarily arose from the constrained road-following trajectory of the aerial work platform, which limited the available viewpoints during acquisition. This analysis suggests that the primary sources of error are from the geometry of the data acquisition rather than from the proposed method.

Table 4: Jumper line detection accuracy per site.

Sensor	Site	Number of Jumper Lines		Accuracy
		Detected Jumper Lines	Ground Truth	
Multi-modal sensor	GN	17	18	94.4
w/o LiDAR		11		61.1
w/o Camera		16		88.8
Multi-modal sensor	MY	8	8	100.0
w/o LiDAR		4		50.0
w/o Camera		8		100.0
Multi-modal sensor	SY	30	32	93.8
w/o LiDAR		21		65.6
w/o Camera		27		84.4

4.2 Accuracy of 4D Multi-Modal Map

This subsection evaluates the effectiveness of the proposed method for constructing the 4D multi-modal map of this study. Table 5 presents the comparative results of four representative state-of-the-art frameworks: VINS-MONO [65], LVI-SAM [66], R2LIVE [67], and R3LIVE [27]. These methods were compared with the proposed method because they also generate a multi-modal map based on RGB–PCD information. The evaluation employed the three-field dataset to secure consistent environmental conditions with the metrics of absolute trajectory error (ATE) and relative pose error (RPE), because these metrics represent global trajectory accuracy and local pose consistency [68].

In a static scenario of the GN site, the proposed method achieved 0.02 m ATE and 0.45° RPE for constructing a 4D multi-modal map. All methods achieved nearly identical performances because the global drift and local deviation remained negligible. Although quantitative differentiation is inherently limited under static conditions, the results verified that the proposed and other methods consistently ensure accuracy in constructing a 4D multi-modal map. These results suggest that all methods successfully constructed a multi-modal map in static environments, where the performance variations were minimal.

Table 5: RPE and ATE comparison of RGB–PCD-based SLAM methods.

Method	Accuracy					
	ATE [m]			RPE [degree]		
	GN	MY	SY	GN	MY	SY
VINS-MONO [65]	0.06	0.15	0.38	0.61	1.08	1.24
LVI-SAM [66]	0.02	0.14	0.26	0.47	0.99	1.10
R2LIVE [67]	0.04	0.14	0.29	0.50	1.09	1.13
R3LIVE [27]	0.02	0.09	0.24	0.45	0.97	1.09
Proposed method	0.02	0.08	0.15	0.45	0.94	1.05

In a short-range driving scenario at the MY site, the proposed method achieved the lowest ATE of 0.08 and RPE of 0.94. Specifically, VINS-MONO, which relies solely on monocular imagery and inertial sensing, recorded 0.15 m ATE and 1.08° RPE because of accumulated drift in texture-sparse regions. An inherent limitation of monocular odometry is that image-based measurements cannot achieve the geometric precision of LiDAR, resulting in degraded accuracy in complex environments [65]. The comparison of ATE and RPE quantitatively reveals this limitation of VINS-MONO, where the higher ATE clearly indicates accumulated trajectory drift while the larger RPE reflects limited local pose consistency. In contrast, LVI-SAM and R2LIVE achieved lower values in both metrics than those by the proposed method. This observation can be explained by the fact that LVI-SAM and R2LIVE utilize a loosely coupled strategy in which LIO and VIO estimate independently and are fused only in the optimization phase. This approach has a limitation in ensuring feature-level constraints across multiple modalities, thereby producing large pose errors under dynamic driving conditions [14]. Note that the proposed method addresses tightly coupled fusion in which LIO, VIO, and GNSS measurements are jointly integrated during state propagation. This tightly coupled integration enables local pose refinement and global drift suppression to act simultaneously, leading to a 47% reduction in trajectory error compared with VINS-MONO and 13% reduction in pose error compared with LVI-SAM. Interestingly, R3LIVE achieved 0.09 m ATE and 0.97° RPE, suggesting that these values from two metrics are very close to those from the proposed method. This high global trajectory accuracy and local pose consistency can be attributed to the utilization of tightly coupled integration in both the proposed method and R3LIVE [27]. However, the proposed method demonstrated superior robustness in complex structural environments because state propagation is enhanced by IMU–GNSS integration, which constrains long-term drift and provides globally consistent motion priors for LiDAR–camera fusion. By contrast, R3LIVE relies only on IMU-based state propagation, resulting in gradual trajectory deviation in extended or irregular scenes. These results demonstrate that tightly coupled integration provides high reliability in terms of trajectory accuracy and pose consistency compared to loosely coupled integration under short-range driving conditions.

In a long-range dynamic scenario of the SY site, the proposed method also achieves the lowest ATE of 0.15 m and RPE of 1.05°, whereas other methods indicate ATE ranging from 0.24 to 0.38 m and RPE ranging from 1.09° to 1.24°. This analysis clearly suggests that the proposed method is superior to other methods. In particular, the proposed method outperformed R3LIVE, which exhibited a performance similar to that in the other two scenarios. Specifically, the proposed method reduces the trajectory error by 38% and pose error by 4%. This difference originates from the GNSS correction used in the proposed method. Note that both methods employ tightly coupled integration of LIO and VIO, but the proposed method employs GNSS correction. Specifically, GNSS measurements are integrated directly into the state propagation process in

the proposed method, whereas R3LIVE relies on IMU-only state propagation, which inevitably accumulates error over extended distances [27]. Note that the errors at the SY site were larger than those at the GN and MY sites because the longer trajectory of the SY site intensified the accumulated drift and local deviation compared with those of the GN and MY sites. Hence, the effect of GNSS correction was particularly evident at the SY site. The analysis of the SY site highlights the advantages of GNSS-assisted state propagation. Specifically, GNSS-assisted correction preserves trajectory accuracy and pose consistency under long-range dynamic driving conditions. LVI-SAM and R2LIVE, which rely on loosely coupled integration, also show larger errors of 0.26–0.29 m in ATE and 1.10–1.13° in RPE, confirming that loosely coupled integration cannot constrain drift accumulation under long-range dynamic conditions.

4.3 Superiority of SDFNet

This subsection evaluates the performance and generalizability of SDFNet through the KITTI-3D benchmark and the field dataset in comparison with representative state-of-the-art frameworks. These neural networks include BtcDet [45], PV-RCNN [46], GLENet [69], and EPNet [70] to encompass both LiDAR-only and LiDAR–image fusion architectures. The quantitative metrics of AP and mAP were used to assess the capability of each neural network to localize and classify instances under different levels of scene complexity.

First, the detection accuracy of the NNs on the field dataset was analyzed to evaluate the accuracy of the SDFNet in our application (Table 6). Remarkably, the SDFNet achieved 77.4% AP for telegraph pole detection, securing the highest accuracy among all compared networks. This result verifies the robustness of the proposed architecture in maintaining stable and reliable detection performance across diverse field conditions. This superior performance is attributed to the pseudo stream architecture, which jointly encodes geometric and semantic representations before integration, thereby ensuring coherent multi-modal alignment and stable detection. LiDAR-only networks, including BtcDet [45], PV-RCNN [46], and GLENet [69] show limited accuracy, with AP values of 73.4%, 72.8%, and 74.5%, respectively, suggesting that extracting features from both 3D PCD and images would be effective for the detection of telegraph poles. However, EPNet [70] fails to outperform LiDAR-only methods. This observation can be explained by the fact that the direct concatenation of image and LiDAR features introduces spatial–semantic misalignment, resulting in the lowest accuracy. This structural limitation degrades the coherence of feature representations, thereby confirming the necessity of geometry-aware feature integration for multi-modal fusion. Specifically, the pseudo stream in SDFNet refines the correspondence between RGB and PCD features through geometry-aware encoding, enabling stable extraction of spatially matched image features compared with the direct integration strategy of EPNet [70].

Table 6: Prediction accuracy on telegraph pole with tree labeling.

Sensor	Network	AP		
		w/o Tree Labeled	w/Tree Labeled	
		Telegraph Pole	Telegraph Pole	Tree
LiDAR	BtcDet [45]	73.4	84.6	60.4
	PV-RCNN [46]	72.8	81.8	63.3
	GLENet [69]	74.5	81.2	66.7
LiDAR & Optical	EPNet [70]	70.3	79.9	55.7
	SDFNet (proposed)	77.4	87.6	69.6

Second, the application of tree labeling was analyzed because this strategy substantially improved the detection accuracy of all networks. Remarkably, the accuracy of SDFNet increased from 77.4% to 87.6% for detecting telegraph poles. Similar improvements were observed among all compared NNs, confirming that tree labeling enhances the capability of NNs to distinguish telegraph poles from adjacent vegetation, leading to a reduction in false detections near vegetated regions. Specifically, the annotated tree instances provide additional contextual information during feature learning, enabling refined spatial attention around pole regions and suppressing irrelevant geometric responses from nearby vegetation. Note that, although the detection accuracy for trees remains relatively low at 69.6%, this limitation does not affect the identification of telegraph poles; rather, it provides auxiliary supervision that stabilizes the multimodal fusion process and improves localization accuracy. The labeling of trees contributes positively to telegraph pole detection by guiding the network to distinguish the target objects of telegraph poles from background vegetation, leading to a reduction in false positive vegetated areas. Furthermore, the inclusion of tree labels provided additional contextual constraints during feature learning, enabling SDFNet to refine spatial attention around pole regions and suppress irrelevant geometric responses from adjacent trees. This auxiliary supervision enhances the discriminative capability of SDFNet in partially occluded environments, leading to improvement in classification and localization. Hence, the effectiveness of this tree labeling strategy demonstrates that accurate background annotation serves not only as a supplementary dataset component but also as a critical factor for stabilizing the multi-modal fusion process in SDFNet.

Finally, the detection accuracy of the KITTI-3D benchmark dataset was analyzed. This analysis aimed to evaluate the overall detection accuracy and general applicability of SDFNet compared with standard 3D object detection tasks (Table 7). Remarkably, SDFNet achieved the highest overall detection accuracy among all neural networks for 3D object detection. Specifically, SDFNet showed an overall mAP of 86.3%. This superior performance of SDFNet can be attributed to the structural advantages of the pseudo stream fusion architecture. In particular, LiDAR-only networks, including BtcDet [45], PV-RCNN [46], and GLENet [69], rely on geometric information, which restricts their capability to recognize objects with low density and geometric contrast. Hence, these neural networks achieve high accuracy for vehicles but exhibit noticeable degradation for cyclists and pedestrians under complex backgrounds. EPNet shows better accuracy than these LiDAR-only networks because it employs a LiDAR–image fusion strategy based on the direct concatenation of encoded features [70]. However, this approach often fails to ensure spatial–semantic correspondence between multiple modalities, resulting in unstable learning and inconsistent accuracy across difficulty levels. SDFNet overcomes these limitations by employing pseudo stream fusion, which jointly encodes geometric and semantic representations before integrating each feature. Joint encoding preserves the spatial correspondence between LiDAR and image features, leading to coherent cross-modal alignment and balanced detection accuracy across the car, cyclist, and pedestrian classes. This distinct architecture enables SDFNet to exploit complementary information from both modalities while mitigating the spatial–semantic misalignment that frequently degrades the performance of conventional fusion networks, including EPNet [70]. Interestingly, BtcDet attained slightly higher accuracy than SDFNet at the hard level. This outcome may be associated with the simple and vertically oriented geometry of pedestrian instances, which are effectively reconstructed by the behind-the-curtain method [45]. This tendency is also attributed to the geometric regularity of symmetric or planar structures, where stable spatial consistency enables reliable depth completion. However, this method is less effective when object geometry is irregular or asymmetric, such as in complex scenes with diverse shapes and orientations.

In summary, SDFNet preserves stable accuracy even under diverse conditions, confirming that the proposed architecture ensures uniform generalization and mitigates class-specific bias across various levels

of scene complexity. Specifically, the pseudo-stream fusion architecture effectively preserves the geometric–semantic correspondence between LiDAR and image features, enabling consistent generalization, even in complex field environments. The incorporation of tree labeling also reinforces this stability by enhancing spatial discrimination between telegraph poles and surrounding vegetation.

Table 7: Comparison with state-of-the art on the KITTI-3D benchmark dataset.

Sensor	Network	Difficulty level	AP			mAP	
			Car	Cyclist	Pedestrian		
LiDAR	BtcDet [45]	Easy	94.8	75.3	90.6	86.9	83.9
		Moderate	92.8	69.5	88.5	83.6	
		Hard	89.6	67.0	87.0	81.2	
	PV-RCNN [46]	Easy	93.5	85.7	86.5	88.6	84.6
		Moderate	89.9	80.5	83.5	84.6	
		Hard	87.2	78.0	76.5	80.6	
	GLENet [69]	Easy	84.7	86.2	83.8	84.9	82.1
		Moderate	91.6	79.9	79.9	83.8	
		Hard	88.5	72.1	72.1	77.6	
LiDAR & Optical	EPNet [70]	Easy	94.9	82.0	90.4	89.1	84.9
		Moderate	94.1	80.0	88.3	87.5	
		Hard	88.8	70.0	75.5	78.1	
	SDFNet (proposed)	Easy	97.3	86.4	84.6	89.4	86.3

4.4 Superiority of RoMP-Tax

This subsection evaluates the performance and generalizability of RoMP-Tax in comparison with representative state-of-the-art oriented-object-detection frameworks, including OBBDetection [42], OrientedFormer [71], RHINO [72], GRA [73], and RoMP-T [54]. These neural networks were selected for quantitative comparison because these neural networks represent image-based architecture designed for rotational bounding-box detection. This analysis clearly demonstrates that the combination of texture transformation through LBP and data augmentation using Mixup not only enhances feature representation but also strengthens the overall robustness of RoMP-Tax. The quantitative metrics of precision, recall, AP, and mAP were used to assess the ability of each network to localize and classify rotated instances under different levels of scene complexity.

First, the detection performance on the field dataset was analyzed to evaluate the accuracy and robustness of RoMP-Tax compared to representative rotated object detection networks (Table 8). Remarkably, the proposed neural network, i.e., RoMP-Tax, achieved 93.4% precision, 90.0% recall, and 91.3% mAP, outperforming representative rotated-object detection neural networks by averages of 13.5%, 7.1%, and 9.5%, respectively. This improvement suggests that the hybrid architecture of CNN–Transformer, texture extraction with LBP, and Mixup augmentation effectively combines the local texture representation with global contextual reasoning, ensuring robust and stable rotated-object detection in complex field environments. Networks such as OBBDetection [42], OrientedFormer [71], and GRA [73], which utilize either a CNN or Transformer architecture, restrict their ability to capture both local and global features within the same representation space. RoMP-T [54] also results in insufficient integration between fine-grained texture cues

and contextual dependencies because it employs a hybrid CNN–Transformer architecture but excludes LBP-based texture extraction and Mixup augmentation. Interestingly, RHINO [72] shows a slightly higher recall performance. This observation can be attributed to the Hausdorff distance–based matching cost, which refines the localization precision during training because the metric directly penalizes the maximum boundary deviation between the predicted and ground-truth boxes, thereby enforcing alignment along object edges rather than overall overlap. However, this boundary-oriented optimization tends to increase false-positive responses under complex textures and irregular boundaries, thereby reducing the precision compared with RoMP-Tax and RoMP-T. Note that RoMP-Tax maintains a balanced precision–recall trade-off by integrating LBP-based texture encoding and Mixup augmentation, thereby enhancing local feature distinctiveness while preventing over-adaptation to specific geometric patterns.

Table 8: Comparison with state-of-the-art on field and DOTA-v2.0 datasets.

Dataset	Network	Precision	Recall	mAP
Field dataset	OBBDetection [42]	69.4	75.6	75.0
	OrientedFormer [71]	71.6	79.9	74.9
	RHINO [72]	86.0	91.7	89.8
	GRA [73]	81.6	82.4	82.2
	RoMP-T [54]	90.8	85.1	87.0
	RoMP-Tax (proposed)	93.4	90.0	91.3
DOTA v2.0	OBBDetection [42]	49.3	54.8	52.6
	OrientedFormer [71]	50.4	52.0	50.8
	RHINO [72]	57.6	61.4	59.9
	GRA [73]	53.3	57.9	55.1
	RoMP-T [54]	59.3	56.4	58.5
	RoMP-Tax (proposed)	61.7	59.9	61.4

Furthermore, the generalizability of RoMP-Tax was evaluated using the DOTA v2.0 dataset [60], which contains various aerial scenes with diverse orientations and scales. Similar to the detection performance on the field dataset, RoMP-Tax showed superior performance compared with other representative neural networks for rotated-object detection. Specifically, the RoMP-Tax achieved a precision of 61.7%, a recall of 59.9%, and mAP of 61.4%, corresponding to average gains of 7.7%, 3.4%, and 6.0%, respectively, over other representative rotated-object detection networks. These results confirm the strong generalizability and broad applicability of RoMP-Tax beyond the field dataset. RHINO also exhibited a relatively high recall value on the DOTAv2.0 dataset, which originated from its Hausdorff distance–based matching loss. However, RoMP-Tax achieved a balanced improvement in both precision and recall by enhancing texture representation through LBP extraction and improving robustness via Mixup augmentation. Note that both the Transformer and CNN, including OBBDetection [42], OrientedFormer [71], and GRA [73], primarily rely on global feature aggregation and orientation regression, which limits the ability of neural networks to extract localized features and correlation compared with the hybrid CNN–Transformer hybrid architecture employed in RoMP-Tax. This integrated design enables RoMP-Tax to maintain a high recall even for small or highly rotated objects, thereby ensuring stable precision and robust generalization across large-scale aerial environments.

To elucidate the contribution of texture extraction and Mixup data augmentation to performance enhancement, the field dataset from the GN, MY, and SY sites was utilized for RoMP-Tax (Table 9).

The baseline configuration recorded an AP of 87.0% without LBP and Mixup. Mixup augmentation increased the AP to 90.8% because this method stabilizes the training process, thereby improving generalization under limited data conditions. LBP-based texture extraction also increased the AP to 90.5% because this method enhances the discrimination of texture features between background and target regions. Remarkably, the integration of LBP and Mixup with the hybrid CNN–Transformer architecture achieved the highest AP of 91.3%, verifying that feature- and data-level enhancements provide complementary effects. Note that this synergy reinforces the robustness of the network against illumination variations and background clutter, ensuring consistent detection performance across diverse scenes. When compared with the results on the DOTA v2.0 dataset presented in Table 8, this improvement indicates that the effect of LBP-based texture encoding and Mixup augmentation might become more pronounced when the available training data are limited, demonstrating the scalability of RoMP-Tax across different data densities.

Table 9: Effectiveness of the texture extraction and Mixup augmentation.

Network	Texture	Augmentation	mAP
RoMP-Tax	w/o LBP	w/o Mixup	87.0
		w/Mixup	90.8
	w/LBP	w/o Mixup	90.5
		w/Mixup	91.3

In summary, RoMP-Tax achieved stable and superior detection performance across both field and large-scale benchmark datasets, verifying that the proposed neural network architecture ensures consistent generalization under diverse scene conditions. The integration of LBP-based texture encoding and Mixup augmentation enhances feature robustness and stabilizes the learning process, leading to reliable recognition of oriented objects with complex textures and varying illumination. This hybridization further demonstrates greater effectiveness in data-limited environments, confirming that RoMP-Tax maintains a uniform generalizability and scalability, regardless of dataset size or texture complexity.

5 Conclusion

This paper proposed a jumper line detection method to support maintenance planning in live distribution facilities. The proposed framework integrates 4D multi-modal mapping, SDFNet, and RoMP-Tax for field inspection of distribution facilities. The 4D multi-modal map is constructed by fusing LiDAR, camera, IMU, and GNSS measurements to generate an accurate spatial representation of the maintenance region. This mapping process ensures a consistent perception of complex environments, overcoming the limitations of motion distortion and sparse sensor observations in conventional mapping approaches. SDFNet detects utility poles from multi-modal sensor data through geometry, pseudo, and fusion streams that jointly learn geometric and optical features. The detected pole coordinates are aggregated within a global reference frame by employing an iKD-Tree, providing consistent spatial information for subsequent analysis. The RoMP-Tax network detects insulators and isolates jumper lines from localized point-cloud regions. The hybrid CNN–Transformer architecture, enhanced by LBP-based texture encoding and Mixup augmentation, enables the robust recognition of small and highly rotated components under complex textures and variable illumination conditions. Comprehensive experiments conducted on field and aerial datasets demonstrated significant improvements in precision, recall, and mAP, thereby validating the robustness and generalizability of the proposed framework. Although the proposed framework demonstrates high accuracy in detecting and localizing jumper lines, the present study is primarily focused on the perception phase under experimental conditions.

Further validation is required to evaluate the performance of this method within actual live-line maintenance workflows using aerial lift vehicles. The integrated architecture provides a reliable and efficient solution for distribution line inspection and enables optimal bucket positioning for autonomous maintenance operations. Future work will utilize the extracted jumper-line position information to support real-time task planning and collision avoidance during high-voltage proximity operations. Furthermore, the framework will be extended to multi-robot cooperation and dynamic state estimation. These advancements will ultimately enable fully autonomous decision-making in large-scale power distribution facility environments.

Acknowledgement: None.

Funding Statement: This research was supported by Developing digital safety measurement to enhance the availability of smart structural monitoring of facilities funded by Korea Research Institute of Standards and Science (KRISS–2025–GP2025-0009), and Development of real-time electrical fire prevention system technology reflecting the characteristics of traditional markets (RS202502634755) by National Fire Agency (NFA, Republic of Korea). Additionally, this research was supported by the Regional Innovation System & Education (RISE) through the Seoul RISE Center, funded by the Ministry of Education (MOE) and the Seoul Metropolitan Government (2026-RISE-01-027-04).

Author Contributions: Joonhyeok Moon: methodology, formal analysis, writing—original draft; Siheon Jeong: investigation & validation; Byeonghyun Lee: software & visualization; Jeik Choi: data curation; Ki-Yong Oh: conceptualization, writing—review & editing, supervision, project administration, funding acquisition. All authors reviewed and approved the final version of the manuscript.

Availability of Data and Materials: The datasets generated and analyzed during the current study are not publicly available due the fact that they constitute an excerpt of research in progress but are available from the corresponding author on reasonable request.

Ethics approval: Not applicable.

Conflicts of Interest: The authors declare no conflicts of interest.

Supplementary Materials: The supplementary material is available online at https://www.techscience.com/doi/10.32604/cmes.2026.081475/s1, References [74,75] are cited in the Supplementary Materials.

Nomenclature

Abbreviation	Full description
2D/3D/4D	Two/Three/Four Dimensional
BEV	Bird's-eye view
BO	Bayesian optimization
BtcDet	Behind-the-curtain detection
CNN	Convolutional neural network
DBSCAN	Density-based spatial clustering of applications with noise
DFOV	Diagonal field of view
Dist-YOLO	Distance-estimation you only look once
ENU	East-north-up
FAST-LIO	Fast LiDAR-inertial odometry
Faster R-CNN	Faster region-based convolutional neural network
GNSS	Global navigation satellite system
ICP	Iterative closest point
ikd-Tree	Incremental k-dimensional tree
LBP	Local binary pattern

LiDAR	Light detection and ranging
LINS	LiDAR-inertial-network-based simultaneous localization and mapping
LIO-SAM	LiDAR inertial odometry via smoothing and mapping
LIO/VIO	LiDAR-inertial odometry/Visual-inertial odometry
LOAM	LiDAR odometry and mapping
LVI-SAM	LiDAR-visual-inertial odometry via smoothing and mapping
M2Det	Multi-level multi-scale detector
MLP	Multi-layer perceptron
MSC	Mean shift clustering
MSML	Multi-scale multi-level
PCD	Point cloud data
PnP	Point-to-plane
PV-RCNN	Point-voxel region-based convolutional neural network
PVT	Pyramid vision transformer
RoMP-Tax	Rotational multi-pyramid transformer with texture and augmentation
SDFNet	Sparse-dense fusion network
SLAM	Simultaneous localization and mapping
YOLO	You only look once

References

1. Ateba BB, Prinsloo JJ, Gawlik R. The significance of electricity supply sustainability to industrial growth in South Africa. Energy Rep. 2019;5(2):1324–38. doi:10.1016/j.egyr.2019.09.041.
2. Jurkiewicz B, Smyrak B. Studies on the evolution of fatigue strength of aluminium wires for overhead line conductors. Materials. 2024;17(11):2537. doi:10.3390/ma17112537.
3. Fogliatto MSS, Caetano HO, Desuó NL, Massignan JAD, Fanucchi RZ, London JBA, et al. Power distribution system interruption duration model using reliability analysis regression. Electr Power Syst Res. 2022;211(7):108193. doi:10.1016/j.epsr.2022.108193.
4. Qiao X, Ming Y, Xu K, Yi N, Sundararajan R. Aging of polymeric insulators under various conditions and environments: another look. Energies. 2022;15(23):8809. doi:10.3390/en15238809.
5. Thomas OO, Chouinard L, Langlois S. Probabilistic fatigue fragility curves for overhead transmission line conductor-clamp assemblies. Front Built Environ. 2022;8:833167. doi:10.3389/fbuil.2022.833167.
6. Cawley JC, Homce GT. Occupational electrical injuries in the United States, 1992–1998, and recommendations for safety research. J Safety Res. 2003;34(3):241–8. doi:10.1016/s0022-4375(03)00028-8.
7. Brenner B, Cawley JC. Occupations most at-risk in fatal overhead power line incidents: using osha data to get a better understanding. In: Proceedings of the 2015 IEEE IAS Electrical Safety Workshop; 2015 Jan 26–30; Louisville, KY, USA. p. 1–6. doi:10.1109/ESW.2015.7094939.
8. Aracil R, Penin LF, Feme M, Jimenez LM, Barrientos A, Santamaria A, et al. ROBTET: A new teleoperated system for live-line maintenance. In: Proceedings of the ESMO'95-1995 IEEE 7th International Conference on Transmission and Distribution Construction, Operation and Live-Line Maintenance; 1995 Oct 29–Nov 3; Columbus, OH, USA. p. 205–11.
9. Chatzargyros G, Papakonstantinou A, Kotoula V, Stimoniaris D, Tsiamitros D. UAV inspections of power transmission networks with AI technology: A case study of Lesvos island in Greece. Energies. 2024;17(14):3518. doi:10.3390/en17143518.
10. Liu J, Zhao Z, Ji J, Hu M. Research and application of wireless sensor network technology in power transmission and distribution system. Intell Converged Netw. 2020;1(2):199–220. doi:10.23919/ICN.2020.0016.
11. Ollero A, Suarez A, Papaioannidis C, Pitas I, Marredo JM, Hoang VD, et al. Multi-aerial robotic system for power line inspection and maintenance: comparative analysis from the AERIAL-CORE final experiments. IEEE Trans Field Robot. 2025;2(1):549–73. doi:10.1109/TFR.2025.3586991.

12. Katrasnik J, Pernus F, Likar B. A survey of mobile robots for distribution power line inspection. IEEE Trans Power Deliv. 2010;25(1):485–93. doi:10.1109/TPWRD.2009.2035427.
13. Cristóvão MP, Portugal D, Carvalho AE, Ferreira JF. A LiDAR-camera-inertial-GNSS apparatus for 3D multimodal dataset collection in woodland scenarios. Sensors. 2023;23(15):6676. doi:10.3390/s23156676.
14. Fan Z, Zhang L, Wang X, Shen Y, Deng F. LiDAR, IMU, and camera fusion for simultaneous localization and mapping: a systematic review. Artif Intell Rev. 2025;58(6):174. doi:10.1007/s10462-025-11187-w.
15. Ali Siddiqui Z, Park U. A drone based transmission line components inspection system with deep learning technique. Energies. 2020;13(13):3348. doi:10.3390/en13133348.
16. Kaartinen E, Dunphy K, Sadhu A. LiDAR-based structural health monitoring: applications in civil infrastructure systems. Sensors. 2022;22(12):4610. doi:10.3390/s22124610.
17. Halder S, Afsari K. Robots in inspection and monitoring of buildings and infrastructure: A systematic review. Appl Sci. 2023;13(4):2304. doi:10.3390/app13042304.
18. Ren Z, Skjetne R, Jiang Z, Gao Z, Verma AS. Integrated GNSS/IMU hub motion estimator for offshore wind turbine blade installation. Mech Syst Signal Process. 2019;123(4):222–43. doi:10.1016/j.ymssp.2019.01.008.
19. Tiozzo Fasiolo D, Scalera L, Maset E. Comparing LiDAR and IMU-based SLAM approaches for 3D robotic mapping. Robotica. 2023;41(9):2588–604. doi:10.1017/s026357472300053x.
20. Zou Z, Yuan C, Xu W, Li H, Zhou S, Xue K, et al. LTA-OM: long-term association LiDAR-IMU odometry and mapping. J Field Robot. 2024;41(7):2455–74. doi:10.1002/rob.22337.
21. Zhang J, Singh S. LOAM: Lidar odometry and mapping in real-time.. In: Robotics: science and systems X. College Station, TX, USA: Robotics: Science and Systems Foundation; 2014. p. 1–9. doi:10.15607/rss.2014.x.007.
22. Shan T, Englot B, Meyers D, Wang W, Ratti C, Rus D. LIO-SAM: tightly-coupled lidar inertial odometry via smoothing and mapping. In: Proceedings of the 2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS); 2020 Oct 24–2021 Jan 24; Las Vegas, NV, USA. p. 5135–42. doi:10.1109/iros45743.2020.9341176.
23. Li Y, Yang S, Xiu X, Miao Z. A spatiotemporal calibration algorithm for IMU-LiDAR navigation system based on similarity of motion trajectories. Sensors. 2022;22(19):7637. doi:10.3390/s22197637.
24. Yin S, Xie D, Fu Y, Wang Z, Zhong R. Uncontrolled two-step iterative calibration algorithm for lidar-IMU system. Sensors. 2023;23(6):3119. doi:10.3390/s23063119.
25. Xu W, Zhang F. FAST-LIO: a fast, robust LiDAR-inertial odometry package by tightly-coupled iterated Kalman filter. IEEE Robot Autom Lett. 2021;6(2):3317–24. doi:10.1109/LRA.2021.3064227.
26. Qin C, Ye H, Pranata CE, Han J, Zhang S, Liu M. LINS: A lidar-inertial state estimator for robust and efficient navigation. In: 2020 IEEE International Conference on Robotics and Automation (ICRA); 2020 May 31–Aug 31; Paris, France. p. 8899–906. doi:10.1109/icra40945.2020.9197567.
27. Lin J, Zhang F. R3LIVE: a robust, real-time, RGB-colored, LiDAR-inertial-visual tightly-coupled state estimation and mapping package. In: Proceedings of the 2022 International Conference on Robotics and Automation (ICRA); 2022 May 23–27; Philadelphia, PA, USA. p. 10672–8. doi:10.1109/ICRA46639.2022.9811935.
28. Li Q, Ma Y, He F, Xi S, Xu J. Bionic vision-based intelligent power line inspection system. Comput Math Meth Med. 2017;2017(1):4964287–13. doi:10.1155/2017/4964287.
29. Rahman EU, Zhang Y, Ahmad S, Ahmad HI, Jobaer S. Autonomous vision-based primary distribution systems porcelain insulators inspection using UAVs. Sensors. 2021;21(3):974. doi:10.3390/s21030974.
30. Santos T, Cunha T, Dias A, Moreira AP, Almeida J. UAV visual and thermographic power line detection using deep learning. Sensors. 2024;24(17):5678. doi:10.3390/s24175678.
31. Tomaszewski M, Gasz R, Osuchowski J. Detection of power line insulators in digital images based on the transformed colour intensity profiles. Sensors. 2023;23(6):3343. doi:10.3390/s23063343.
32. Ren S, He K, Girshick R, Sun J. Faster R-CNN: Towards real-time object detection with region proposal networks. IEEE Trans Pattern Anal Mach Intell. 2017;39(6):1137–49. doi:10.1109/tpami.2016.2577031.
33. Zhao Q, Sheng T, Wang Y, Tang Z, Chen Y, Cai L, et al. M2Det: A single-shot object detector based on multi-level feature pyramid network. Proc AAAI Conf Artif Intell. 2019;33(1):9259–66. doi:10.1609/aaai.v33i01.33019259.
34. Bochkovskiy A, Wang CY, Liao HM. YOLOv4: optimal speed and accuracy of object detection. arXiv:2004.10934. 2020.

35. Redmon J, Divvala S, Girshick R, Farhadi A. You only look once: unified, real-time object detection. In: Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR); 2016 Jun 27–30; Las Vegas, NV, USA. p. 779–88. doi:10.1109/CVPR.2016.91.
36. Redmon J, Farhadi A. YOLOv3: An incremental improvement. arXiv:1804.02767. 2018.
37. Redmon J, Farhadi A. YOLO9000: better, faster, stronger. arXiv:1612.08242. 2016.
38. Kim D, Kim S, Jeong S, Ham JW, Son S, Moon J, et al. Rotational multipyramid network with bounding-box transformation for object detection. Int J Intell Syst. 2021;36(9):5307–38. doi:10.1002/int.22513.
39. Xie X, Cheng G, Wang J, Yao X, Han J. Oriented R-CNN for object detection. In: IEEE/CVF International Conference on Computer Vision (ICCV); 2021 Oct 10–17. Montreal, QC, Canada. p. 3500–9. doi:10.1109/iccv48922.2021.00350.
40. Lin TY, Goyal P, Girshick R, He K, Dollár P. Focal loss for dense object detection. In: Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV); 2017 Oct 22–29; Venice, Italy. p. 2999–3007. doi:10.1109/ICCV.2017.324.
41. Xu Y, Fu M, Wang Q, Wang Y, Chen K, Xia GS, et al. Gliding vertex on the horizontal bounding box for multi-oriented object detection. IEEE Trans Pattern Anal Mach Intell. 2021;43(4):1452–9. doi:10.1109/TPAMI.2020.2974745.
42. Yang X, Yan J. On the arbitrary-oriented object detection: Classification based approaches revisited. Int J Comput Vis. 2022;130(5):1340–65. doi:10.1007/s11263-022-01593-w.
43. Yang X, Yang X, Yang J, Ming Q, Wang W, Tian Q, et al. Learning high-precision bounding box for rotated object detection via kullback-leibler divergence. arXiv:2106.01883. 2021.
44. Vajgl M, Hurtik P, Nejezchleba T. Dist-YOLO: fast object detection with distance estimation. Appl Sci. 2022;12(3):1354. doi:10.3390/app12031354.
45. Xu Q, Zhong Y, Neumann U. Behind the curtain: Learning occluded shapes for 3D object detection. Proc AAAI Conf Artif Intell. 2022;36(3):2893–901. doi:10.1609/aaai.v36i3.20194.
46. Shi S, Guo C, Jiang L, Wang Z, Shi J, Wang X, et al. PV-RCNN: point-voxel feature set abstraction for 3D object detection. In: Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2020 Jun 13–19; Seattle, WA, USA. p. 10526–35. doi:10.1109/cvpr42600.2020.01054.
47. Charles RQ, Hao S, Mo K, Guibas LJ. PointNet: deep learning on point sets for 3D classification and segmentation. In: Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR); 2017 Jul 21–26; Honolulu, HI, USA. p. 77–85. doi:10.1109/CVPR.2017.16.
48. Liu Z, Tang H. Learning sparse geometric features for building segmentation from low-resolution remote-sensing images. Remote Sens. 2023;15(7):1741. doi:10.3390/rs15071741.
49. Cai Y, Xu W, Zhang F. Ikd-tree: an incremental K-D tree for robotic applications. arXiv:2102.10808. 2021.
50. Ester M, Kriegel HP, Sander J, Xu X. A density-based algorithm for discovering clusters in large spatial databases with noise. In: Proceedings of the KDD-96; 1996 Aug 2–4; Portland, OR, USA. p. 226–31.
51. Besl PJ, McKay ND. A method for registration of 3-D shapes. IEEE Trans Pattern Anal Mach Intell. 1992;14(2):239–56. doi:10.1109/34.121791.
52. Yu J, Fang H, Zhang X, Wu W, He Y. Tightly coupled GNSS/IMU/vision integrated system for positioning in agricultural scenarios. Comput Electron Agric. 2025;239(105):110478. doi:10.1016/j.compag.2025.110478.
53. Wu X, Peng L, Yang H, Xie L, Huang C, Deng C, et al. Sparse fuse dense: towards high quality 3D detection with depth completion. In: Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2022 Jun 18–24; New Orleans, LA, USA. p. 5408–17. doi:10.1109/CVPR52688.2022.00534.
54. Moon J, Jeon M, Jeong S, Oh KY. RoMP-transformer: rotational bounding box with multi-level feature pyramid transformer for object detection. Pattern Recognit. 2024;147(2):110067. doi:10.1016/j.patcog.2023.110067.
55. Kalra A, Stoppi G, Brown B, Agarwal R, Kadambi A. Towards rotation invariance in object detection. In: Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision (ICCV); 2021 Oct 10–17; Montreal, QC, Canada. p. 3510–20. doi:10.1109/ICCV48922.2021.00351.

56. Murrugarra-Llerena J, Kirsten L, Jung CR. Can we trust bounding box annotations for object detection? In: Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW); 2022 Jun 19–20; New Orleans, LA, USA. p. 4812–21. doi:10.1109/CVPRW56347.2022.00528.
57. Zhang H, Cisse M, Dauphin YN, Lopez-Paz D. Mixup: beyond empirical risk minimization. arXiv:1710.09412. 2017.
58. Ojala T, Pietikäinen M, Harwood D. A comparative study of texture measures with classification based on featured distributions. Pattern Recognit. 1996;29(1):51–9. doi:10.1016/0031-3203(95)00067-4.
59. Geiger A, Lenz P, Urtasun R. Are we ready for autonomous driving? The KITTI vision benchmark suite. In: Proceedings of the 2012 IEEE Conference on Computer Vision and Pattern Recognition; 2012 Jun 16–21; Providence, RI, USA. p. 3354–61. doi:10.1109/CVPR.2012.6248074.
60. Ding J, Xue N, Xia GS, Bai X, Yang W, Yang MY, et al. Object detection in aerial images: a large-scale benchmark and challenges. IEEE Trans Pattern Anal Mach Intell. 2022;44(11):7778–96. doi:10.1109/TPAMI.2021.3117983.
61. Velodyne Lidar. Alpha prime (VLS-128) datasheet. [cited 2026 Jan 1]. Available from: https://www.mapix.com/wp-content/uploads/2019/11/VelodyneLidar_AlphaPrime_Datasheet.pdf.
62. e-con Systems Inc. 3Cam USB 3.0 camera series datasheet. [cited 2026 Jan 1]. Available from: https://www.e-consystems.com/4k-usb-camera.asp.
63. VectorNav. VN-100 IMU/AHRS datasheet. Dallas, TX, USA: VectorNav Technologies; 2025 [cited 2026 Jan 1]. Available from: https://www.vectornav.com/resources/detail/vn-100-imu-ahrs.
64. u-blox. ZED-F9P module data sheet. Thalwil, Switzerland: u-blox; 2025 [cited 2026 Jan 1]. Available from: https://content.u-blox.com/sites/default/files/documents/ZED-F9P-05B_DataSheet_UBXDOC-963802114-12824.pdf.
65. Qin T, Li P, Shen S. VINS-mono: a robust and versatile monocular visual-inertial state estimator. IEEE Trans Robot. 2018;34(4):1004–20. doi:10.1109/TRO.2018.2853729.
66. Shan T, Englot B, Ratti C, Rus D. LVI-SAM: tightly-coupled lidar-visual-inertial odometry via smoothing and mapping. In: Proceedings of the 2021 IEEE International Conference on Robotics and Automation (ICRA); 2021 May 30–Jun 5; Xi'an, China. p. 5692–8. doi:10.1109/icra48506.2021.9561996.
67. Lin J, Zheng C, Xu W, Zhang F. R $ˆ2$ LIVE: A robust, real-time, LiDAR-inertial-visual tightly-coupled state estimator and mapping. IEEE Robot Autom Lett. 2021;6(4):7469–76. doi:10.1109/LRA.2021.3095515.
68. Sturm J, Engelhard N, Endres F, Burgard W, Cremers D. A benchmark for the evaluation of RGB-D SLAM systems. In: Proceedings of the 2012 IEEE/RSJ International Conference on Intelligent Robots and Systems; 2012 Oct 7–12; Vilamoura-Algarve, Portugal. p. 573–80. doi:10.1109/IROS.2012.6385773.
69. Zhang Y, Zhang Q, Zhu Z, Hou J, Yuan Y. GLENet: Boosting 3d object detectors with generative label uncertainty estimation. Int J Comput Vis. 2023;131(12):3332–52. doi:10.1007/s11263-023-01869-9.
70. Huang T, Liu Z, Chen X, Bai X. EPNet: enhancing point features with image semantics for 3D object detection. In: Computer Vision—ECCV 2020. Cham, Switzerland: Springer International Publishing; 2020. p. 35–52. doi:10.1007/978-3-030-58555-6_3.
71. Zhao J, Ding Z, Zhou Y, Zhu H, Du WL, Yao R, et al. OrientedFormer: an end-to-end transformer-based oriented object detector in remote sensing images. IEEE Trans Geosci Remote Sens. 2024;62:5640816. doi:10.1109/TGRS.2024.3456240.
72. Lee H, Song M, Koo J, Seo J. RHINO: rotated DETR with dynamic denoising via Hungarian matching for oriented object detection. arXiv:2305.07598. 2023.
73. Wang J, Pu Y, Han Y, Guo J, Wang Y, Li X, et al. GRA: detecting oriented objects through group-wise rotating and attention. In: Computer Vision—ECCV 2024. Cham, Switzerland: Springer Nature; 2024. p. 298–315. doi:10.1007/978-3-031-72643-9_18.
74. Snoek J, Larochelle H, Adams RP. Practical Bayesian optimization of machine learning algorithms. Adv Neural Inf Process Syst. 2012;25:2951–9.
75. Alibrahim H, Ludwig SA. Hyperparameter optimization: comparing genetic algorithm against grid search and Bayesian optimization. In: Proceedings of the 2021 IEEE Congress on Evolutionary Computation (CEC); 2021 Jun 28–Jul 1; Virtual. p. 1551–9.