Proceedings of the 2023 Winter Simulation Conference
C. G. Corlu, S. R. Hunter, H. Lam, B. S. Onggo, J. Shortle, and B. Biller, eds.

3D OBJECT DETECTION AND LOCALIZATION WITHIN HEALTHCARE FACILITIES

Da Hu

Department of Civil and Environmental Engineering
Kennesaw State University
655 Arntson Drive
Marietta, GA 30060, USA

Mengjun Wang
Shuai Li

Department of Civil and Environmental Engineering
The University of Tennessee
851 Neyland Drive
Knoxville, TN 37996, USA

ABSTRACT

This study introduces a deep learning-based method for indoor 3D object detection and localization in healthcare facilities. This method incorporates spatial and channel attention mechanisms into the YOLOv5 architecture, ensuring a balance between accuracy and computational efficiency. The network achieves an AP50 of 67.6%, an mAP of 46.7%, and a real-time detection rate with an FPS of 67. Moreover, the study proposes a novel mechanism for estimating the 3D coordinates of detected objects and projecting them onto 3D maps, with an average error of 0.24 m and 0.28 m in the x and y directions, respectively. After being tested and validated with real-world data from a university campus, the proposed method shows promise for improving disinfection efficiency in healthcare facilities by enabling real-time object detection and localization for robot navigation.

1 INTRODUCTION

In healthcare facilities, abundant evidence attests to the role of contaminated environmental surfaces and equipment in transmitting pathogens, giving rise to hospital-acquired infectious disease outbreaks. Consequently, the research focus has been directed towards the involvement of inanimate objects in the vicinity of patients in the transmission dynamics of such diseases. It is now broadly acknowledged that patient-proximate surfaces can function as reservoirs for nosocomial pathogens (Huslage et al. 2010). Healthcare workers often serve as vectors, indirectly or directly transferring infectious pathogens to nearby surfaces. The sanitation of high-touch objects in healthcare facilities, crucial for mitigating fomite transmission, primarily falls under the cleaning staff's purview. Nevertheless, factors such as fatigue may impede the effectiveness of manual cleaning (Carling et al. 2010). Additionally, being in contact with high-touch objects in infectious environments exposes cleaning staff to a heightened risk of infection. Promising advancements have been observed in environmental surface decontamination through the deployment of disinfection robots across various settings.

The considerable potential of robots in environmental surface disinfection has drawn significant research interest. For example, Roelofs et al. (2021) created a UAV-based disinfection system primarily targeting door handles. This specialized approach might overlook other high-touch surfaces, potential sources of pathogens. In previous work, we developed algorithms for object detection (Hu et al. 2023a), contaminated area segmentation (Hu et al. 2020), and material classification (Hu and Li 2022a), designed specifically for disinfection robots in healthcare facilities. However, existing disinfection robots encounter challenges in identifying high-touch objects and estimating their 3D coordinates for efficient disinfection. This is due to two primary issues: a dearth of datasets dedicated to high-touch object recognition in healthcare facilities, and the difficulty in translating RGB image-based object detection results to a 3D map

for robot navigation. Accurate high-touch object identification is essential for disinfection robots to target surfaces with a high transmission risk and administer suitable disinfectant dosages.

To address these challenges, this paper introduces a comprehensive deep learning-based solution designed to detect, classify, and project objects found in healthcare facilities onto a 3D map, significantly enhancing the navigation capabilities of disinfection robots. This research's contributions are both innovative and multi-dimensional. First, we proposed an object recognition method by incorporating spatial and channel attention mechanisms into the highly efficient YOLOv5 architecture. This fusion not only maintains high accuracy but also optimizes computational efficiency, enabling real-time application. Second, we created a unique mapping framework that seamlessly integrates object detection, object coordinate estimation, and object clustering. This synergistic approach effectively projects detected objects onto a 3D map, creating an invaluable tool for robot navigation in complex environments. We substantiated the practical relevance and effectiveness of our proposed method by testing and validating it using real data collected from a university campus building. The results conclusively demonstrate the method's feasibility and its potential for broad application in various practical contexts, especially in the critical realm of healthcare facilities.

2 LITERATURE REVIEW

This section reviews algorithms and datasets for the task of object recognition and localization. Several datasets have been created for object detection, such as Microsoft COCO, Open Images, and PASCAL VOC2007. These datasets, comprising both indoor and outdoor environments, have been widely employed as benchmarks to evaluate the performance of deep learning networks. However, very few datasets have been specifically developed for object recognition in healthcare facilities. In this context, Bashiri et al. (2018) proposed an object classification dataset named MCIndoor20000, which uses images collected from the Marshfield Clinic. This dataset contains a total of 2,055 images with three object categories: doors, stairs, and hospital signs. More recently, Ismail et al. (2020) created an image classification dataset (MYNursingHome) using a total of 37,500 images collected in several nursing homes. This dataset comprises 25 indoor object categories such as basket bins, benches, cabinets, chairs, and wheelchairs. The main drawback of the MCIndoor20000 and MYNursingHome datasets is that surrounding backgrounds and objects were removed for each object category. Consequently, every image contains only one object category, rendering the datasets unsuitable for object detection in cluttered indoor environments.

As computational power has advanced, deep learning methods have become widely used in computer vision tasks, proving effective in various fields including building damage detection (Hu et al. 2023b) and radargram inversion (Hu et al. 2022b). In the realm of object detection in healthcare facilities, significant strides have been made. Vasquez et al. (2017) incorporated a fast region proposal method into a Fast R-CNN network, thereby boosting object detection efficiency and speed. The network showed promise in identifying patients with mobility aids, such as wheelchairs. The object detection results were subsequently refined using a probabilistic estimator for position, velocity, and class, generated through a hidden Markov model. Furthermore, Kinasih et al. (2020) employed a 'you only look once' (YOLO) based object detector, specifically to identify hospital beds. To tackle low-confidence detection, a centroid tracking approach was proposed, involving displacement calculation relative to the object size.

Several studies have integrated Simultaneous Localization and Mapping (SLAM) and object detection to estimate the 3D coordinates of recognized objects. For instance, Liu et al. (2022) proposed a method that integrates ORB-SLAM2 with object detection to estimate an object's position in a 3D map. However, this method necessitates the use of ArUco markers for camera calibration, object size estimation, and pose estimation, thereby limiting its applicability in new indoor environments. Similarly, Rosinol et al. (2021) classified objects by aligning the reconstructed 3D object mesh with known object shapes, which also enabled them to estimate the 3D pose of objects. However, this technique requires a well-reconstructed 3D map with accurate 3D reconstructed object shapes, a condition difficult to meet in real-world applications. The present study aims to address this knowledge gap.

3 METHODOLOGY

In this paper, we introduce a comprehensive four-step methodology for generating a semantic 3D map, as illustrated in Figure 1. Our first step involves developing a deep learning-based network designed to detect and classify objects within RGB images. The second step employs SLAM (Simultaneous Localization and Mapping) to estimate camera pose and generate a corresponding 3D map. The third phase combines camera poses, generated point clouds, and identified objects to estimate the objects' 3D coordinates. Lastly, we utilize a density-based algorithm to cluster objects detected from various camera perspectives, using each cluster's centroid to represent the object's position. The projected objects then construct a semantic 3D map. This 3D semantic map offers potential for improving robot navigation and enhancing information registration.

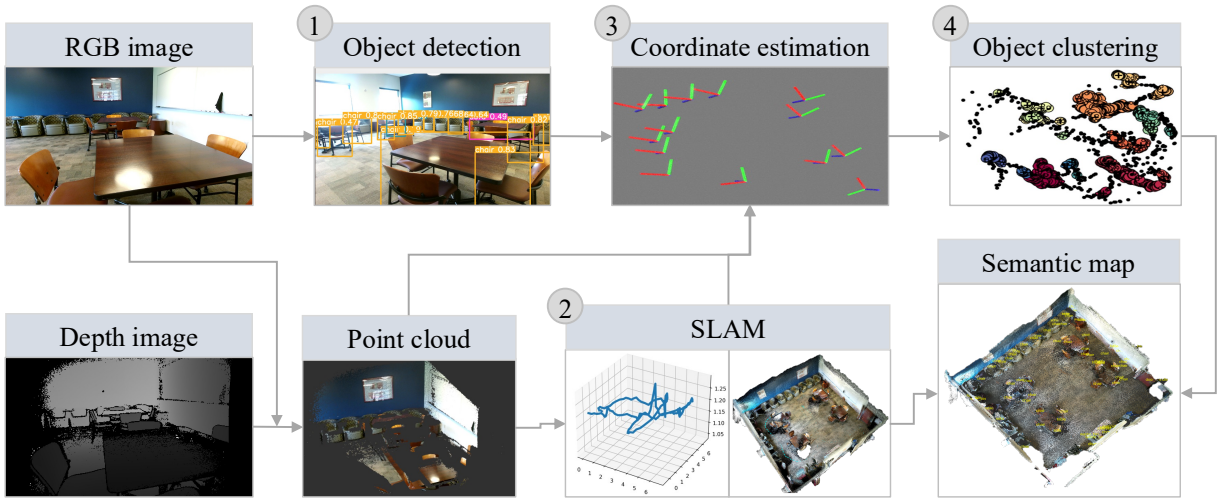


Figure 1: Methodology overview.

3.1 Object Detection Network

This study aims to improve the rapid inference capability of object detection networks by utilizing the YOLOv5s architecture. The YOLOv5s architecture consists of three main components: the backbone network, the detection neck, and three detection heads. The initial stage of the architecture involves image preprocessing using the mosaic method, which is a data augmentation technique designed to improve the network's performance on small objects. The backbone network plays a crucial role in extracting features at various levels from the input images. It is constructed based on the Cross Stage Partial Network (CSPNet) (Wang et al. 2019), which incorporates gradient changes into the feature map from its inception to its conclusion. By doing so, the CSPNet architecture reduces computational costs while preserving the network's inference power. Each CSPNet network comprises of three cascaded convolutional layers with diverse bottlenecks. Furthermore, the backbone layer incorporates the coordinate attention mechanism (Hou et al. 2021), which allows the YOLOv5s network to focus on critical regions with minimal computational expense. The Spatial Pyramid Pooling - Fast (SPPF), which is the final layer of the backbone, concurrently pools on multiple kernel sizes (5, 9, 13) to extract both fine and coarse information. The detection neck, which is based on the Path Aggregation Network (PANet) (Liu et al. 2018), is responsible for enhancing information flow at different levels. The PANet is an improvement over the Feature Pyramid Network (FPN), with the addition of an extra bottom-up pathway. The detection neck obtains feature pyramids that are utilized to identify objects of varying sizes and scales. Comprising four CSPNet blocks, the detection neck generates three feature maps with distinct scales to predict targets of various dimensions. These feature maps are then partitioned into grids, and each grid is assigned three anchors to predict the bounding box for the object. Figure 2 offers a detailed depiction of the adopted network architecture. We

have introduced two enhancements to this model: the addition of an attention mechanism and the replacement of the bounding box regression loss.

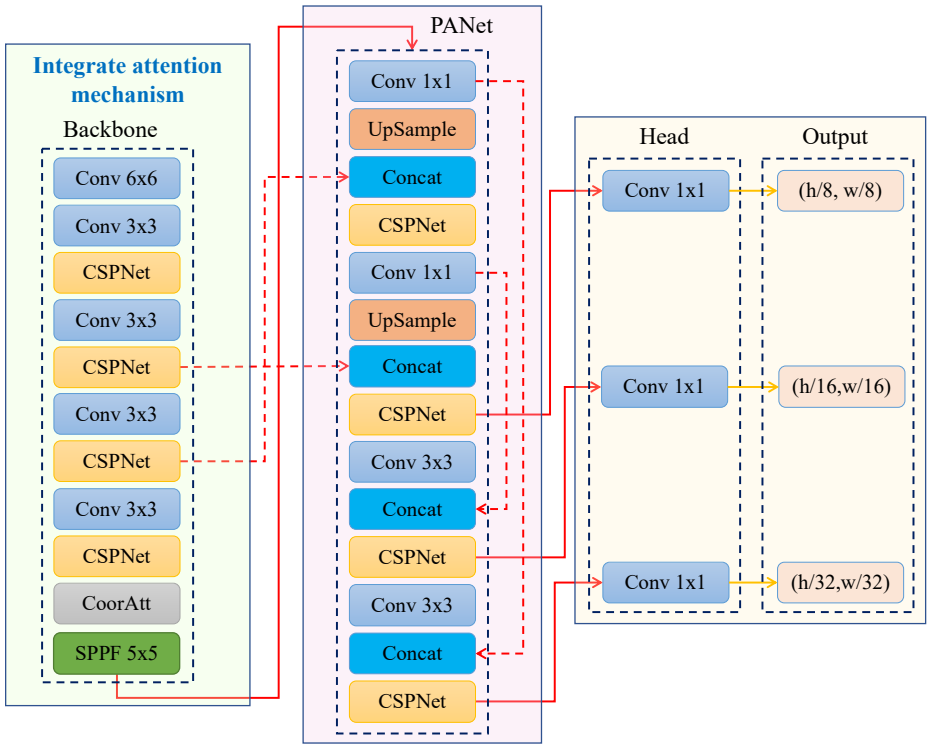


Figure 2: Architecture of the proposed network.

3.2 3D Object Localization

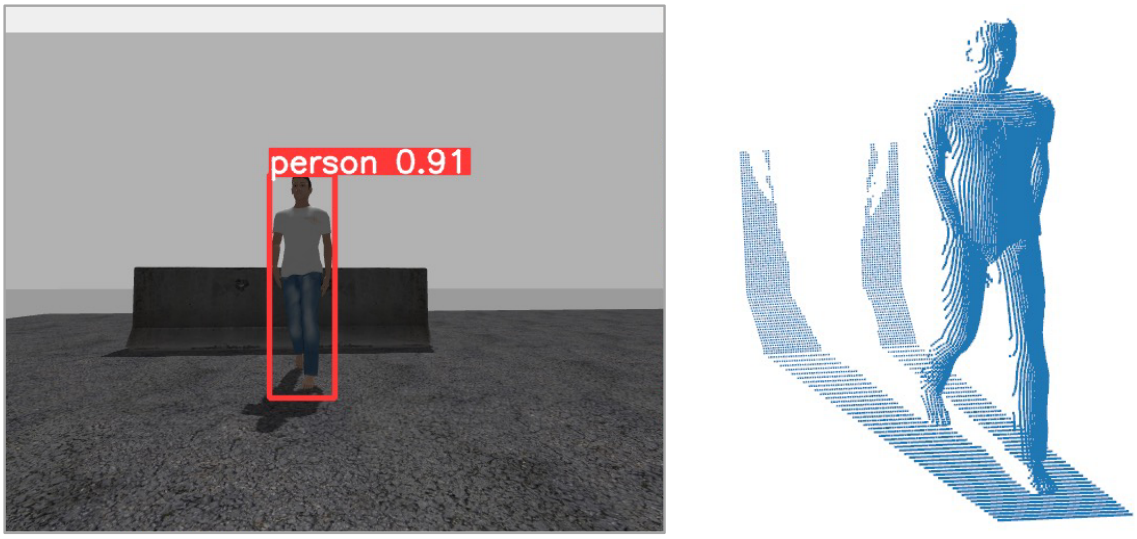
After recognizing objects in 2D RGB images, it is necessary to project the labels to a 3D grid map for robot navigation and disinfection. The classical pinhole camera model (Bradski and Kaehler 2000) is used to calculate the point cloud of the environment using the depth images. The 3D object mapping consists of three steps, namely object coordinate calculation, SLAM, and object clustering.

3.2.1 Object Coordinate Calculation

In order to compute the coordinates of an object detected in an image, the corresponding 3D point cloud must first be identified. For commonly used RGB-D cameras, such as Kinect and RealSense, the depth image is aligned with the RGB image. The object is represented as a bounding box within the image, and the 2D pixel coordinates encompassed by the bounding box are projected onto a 3D point cloud. The object label is converted to cluster point indices under the assumption that 0 represents the background. Consequently, a set of point indices is generated, with each point index signifying an object label. This approach enables the accurate computation of object coordinates based on their respective point cloud representations.

The decomposition operator generates a subset of point clouds with an object label assigned to each detected object. Given that the object is delineated as a bounding box in images, it is inevitable that some points not pertaining to the object might be included in the point cloud subset. These noisy points may originate from the ground surface or other background elements, adversely affecting the accuracy of coordinate estimation. Figure 3 illustrates an example of a clustered point cloud for a human, which includes noisy points. As depicted, the point cloud contains extraneous points from both the ground plane and the

barrier situated behind the individual. It is essential to eliminate these noisy points to ensure precise object position estimation. In this study, two point cloud filters are implemented to remove noisy points, consequently mitigating their impact on coordinate estimation. Initially, the PassThrough filter is applied to directly eliminate points that fail to meet the designated threshold.



(a) (b)

Figure 3: Example point cloud corresponding to the detected person. (a) human detected by the network; and (b) point cloud.

The ground surface points are filtered out using a range threshold in the z-direction dimension. Next, a statistical outlier removal approach is used to remove points that are further away from their neighbors compared to the average for the point cloud. The method can be divided into the following steps.

- Set k, an integer, representing the number of closest points around point P_i,
- Set a standard deviation multiplier α,
- For every point P_i in the 3D point cloud
    - Find the location of k nearest neighbors to point P_i,
    - Compute the average distance d_i from point P_i to its k nearest neighbors,
- Compute the mean μ_d of the distance d_i,
- Compute the standard deviation σ_d of the distance d_i,
- Compute the threshold $T = \mu_d + \alpha \cdot \sigma_d$,
- Eliminate points in the cloud for which the average distance to its k neighbors is at a distance $d > T$.

Figure 4 shows the filtered point cloud using range threshold and statistical outlier removal. Specifically, the range threshold is set from 0.1 to 5 m. The value of k is set to 100, and α is set to 0.5. The ground surface points can be eliminated with the range threshold. The points that come from the background are successfully eliminated, resulting in a clean point cloud for the human. The filtered point cloud is then used to estimate the 3D coordinate of the object.

After filtering, the centroid of point cloud P_c can be estimated using Eq. (1), where N is the total number of points in the filtered point cloud, and (x_i, y_i, z_i) are point coordinates.

$$P_c = \frac{1}{N}\left(\sum_{i=0}^{N} x_i, \sum_{i=0}^{N} y_i, \sum_{i=0}^{N} z_i\right) \quad (1)$$

The 3D bounding box of the point cloud can also be estimated from the filtered point cloud. Note that the z axis of a 3D bounding box is perpendicular to the ground plane. The direction of x and y for the 3D bounding box is estimated using Principal Component Analysis (PCA).

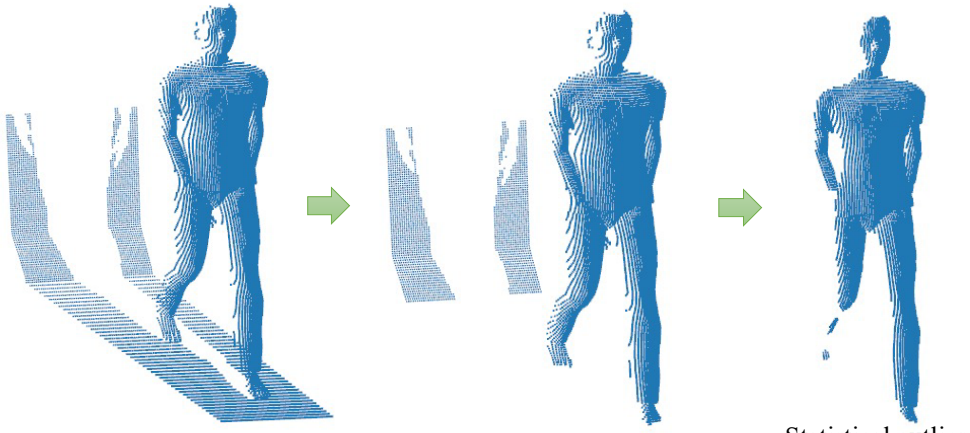


Figure 4: Filtered point cloud.

For PCA estimation, the 3D point cloud is projected onto the ground plane. The covariance matrix is first calculated in Eq. (2), where M ($\mathbf{x}_p$, $\mathbf{y}_p$) are coordinates of projected points on the plane.

$$A = \begin{bmatrix} cov(\mathbf{x}_p, \mathbf{x}_p) & cov(\mathbf{x}_p, \mathbf{y}_p) \\ cov(\mathbf{x}_p, \mathbf{y}_p) & cov(\mathbf{y}_p, \mathbf{y}_p) \end{bmatrix} \tag{2}$$

Given an eigenvalue of λ, an eigenvector V associated with λ for the covariance matrix, A should satisfy Eq. (3). The eigenvector V for A can then be calculated. V_1 and V_2 are directions of the first and second principal components, i.e., directions of x and y, respectively.

$$AV = \lambda V \tag{3}$$

Next, the point cloud M is projected onto principal components using Eq. (4)

$$T = VM \tag{4}$$

3.2.2 Localization and Mapping

In this study, the RTAB-Map SLAM method (Labbé and Michaud 2019), a graph-based SLAM technique, is employed for robot localization and environmental occupancy map generation to facilitate navigation. The map's structure comprises nodes and links. Odometry nodes disseminate odometry data for the estimation of robotic poses. Visual odometry, derived from ORB-SLAM2 (Mur-Artal and Tardós 2017), serves as the odometry input due to its rapidity and precision. The Short-term Memory (STM) module constructs nodes to store odometry and RGB-D images, in addition to computing other information such as visual features and local occupancy grids. A weighting mechanism is implemented to determine the transfer of nodes from Working Memory (WM) to Long-term Memory (LTM), thereby constraining the WM size and reducing graph update time. When loop closure is detected, nodes in the LTM can be reintegrated into the WM. Links store transformation information between two nodes, with neighbor and loop closure links functioning as constraints for graph optimization and odometry drift reduction. The Bag of Words approach

(Kejriwal et al. 2016) is utilized for loop closure detection. Visual features extracted from local feature descriptors, such as ORB (Rublee et al. 2011), are quantized into a vocabulary for expedited comparison. RTAB-Map outputs, including camera pose and 2D occupancy grid, are employed for semantic mapping and robot navigation. The rtabmap-ros package is accessible in ROS, facilitating seamless integration with autonomous robots for this application.

3.2.3 Object Clustering

As the camera is in continuous motion, objects may be detected from varying perspectives, necessitating the clustering of object detection results across multiple viewpoints. Initially, 3D coordinates for detected objects are estimated across different camera views. Given its ability to identify clusters of diverse shapes and sizes without predefining the number of clusters, its efficient handling of outliers, and reduced computational demands, the DBSCAN algorithm is favored over other clustering methods like K-means (Fuchs and Höpken 2022). DBSCAN requires two parameters: the maximum distance epsilon (ε) and the minimum number of points (minPts) constituting a cluster. The algorithm commences by selecting an arbitrary point within the point cloud and retrieving points located within the ε distance. If the number of points exceeds minPts, a cluster is established; otherwise, the point is considered noise. When a point is designated as part of a cluster, its neighboring points within the ε distance are also incorporated into the cluster. The cluster persists in adding new points until no further points are found within the ε distance. Subsequently, a new unvisited point is chosen, and the same procedure is executed to identify clusters or noise. The DBSCAN process is delineated in pseudocode in Figure 5.

```
DBSCAN(DB, distFunc, eps, minPts) {
    C := 0
    for each point P in database DB {
        if label(P) ≠ undefined then continue
        Neighbors N := RangeQuery(DB, distFunc, P, eps)
        if |N| < minPts then {
            label(P) := Noise
            continue
        }
        C := C + 1
        label(P) := C
        SeedSet S := N \ {P}
        for each point Q in S {
            if label(Q) = Noise then label(Q) := C
            if label(Q) ≠ undefined then continue
            label(Q) := C
            Neighbors N := RangeQuery(DB, distFunc, Q, eps)
            if |N| ≥ minPts then {
                S := S ∪ N
            }
        }
    }
}
```

Figure 5: Pseudocode of DBSCAN algorithm (adapted from Schubert et al. (2017)).

4 EXPERIMENT AND RESULTS

4.1 Results on Object Detection

The network is trained on a workstation running Ubuntu 16.04 with dual Intel Xeon Gold 4114 CPU, 128 GB RAM, and NVIDIA RTX A6000. The Stochastic Gradient Descent (SGD) optimizer is used to train the network. The network is trained for a total of 300 epochs. The Hospital Indoor Object Detection (HIOD)

dataset (Hu et al. 2023a) is selected for network evaluation. The dataset is randomly split into a training set (80%), and a validation set (20%). The images are resized to 640 × 640. The early stopping technique is used to avoid the overfitting problem. Specifically, the network stops training when the loss value does not decrease for 20 epochs. The model with the highest performance on the validation set is saved for performance analysis.

The results demonstrate that the proposed method achieves significant performance metrics, with an AP_{50} of 67.6% and a mAP of 46.7% on the validation set of the HIOD dataset. The effectiveness of the alpha-IoU and coordinate attention mechanisms is evaluated by systematically incorporating each component into the baseline model and examining its impact on the model's performance. Table 1 displays the results for the validation set of the HIOD dataset, with the mAP serving as the primary performance evaluation metric for the network. The findings indicate that the baseline network exhibits a performance improvement of 0.6% when augmented by the alpha-IoU mechanism, signifying its positive contribution to the model. Likewise, the coordinate attention module demonstrates its efficacy by enhancing the baseline performance by an equivalent margin of 0.6%. A synergistic integration of both the alpha-IoU and coordinate attention mechanisms achieves the most optimal performance, resulting in a cumulative improvement of 1.7%. This considerable enhancement emphasizes the effectiveness of the proposed method in detecting and classifying building damage, underscoring its potential for practical applications in assessing structural integrity and informing subsequent interventions. Figure 6 illustrates example results of object detection in the validation set of the HIOD dataset. The results indicate that the proposed method can accurately detect and classify objects in healthcare facilities.

Table 1: Ablation study.

Model	alpha-IoU	Coordinate attention	mAP (%)
YOLOv5s	-	-	45.0
	✓	-	45.6
	-	✓	45.6
Proposed	✓	✓	**46.7**

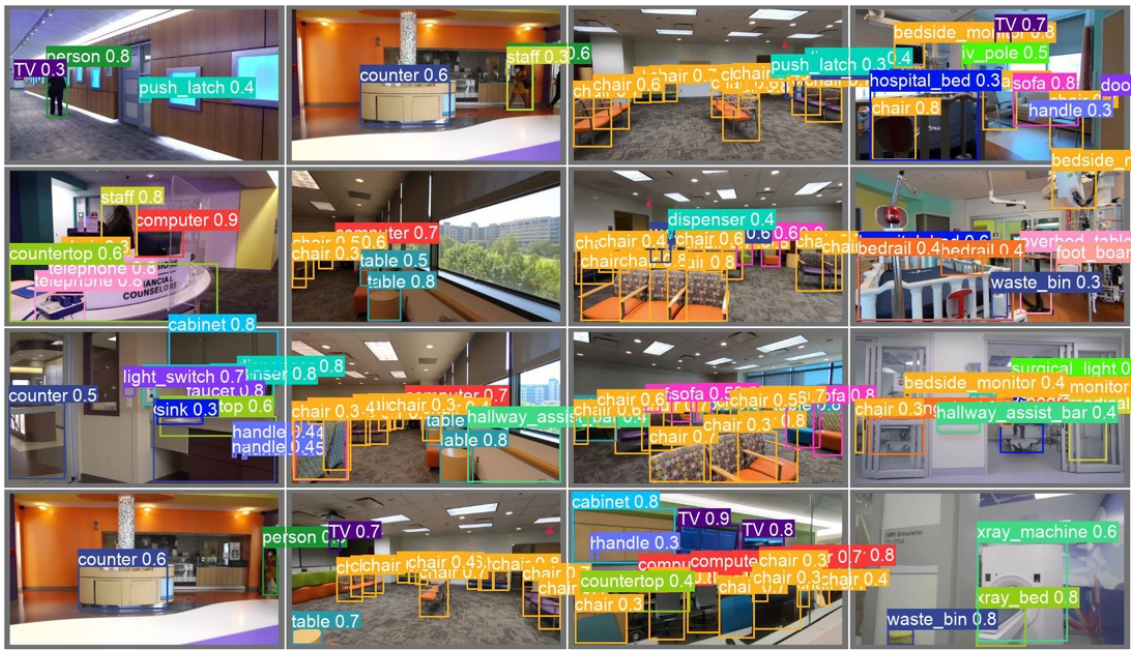


Figure 6: Examples of model prediction results.

4.2 Results on Object Localization

To assess the accuracy of object mapping, a robot simulation platform was developed to emulate indoor environment reconstruction. The platform operates on a laptop running Ubuntu 18.04, with the Robot Operating System (ROS) distribution being Melodic and the Gazebo version being 9. Within the simulation platform, the robot is equipped with an LMS1xx 2D laser and a RealSense RGB-D camera. Figure 7(a) illustrates the experimental setup, in which six individuals are randomly situated within the indoor environment. The human object is selected for evaluation purposes, as the deep learning network, trained using real images, may experience compromised performance in a simulated environment. The human features in the virtual environment closely resemble those in the real world, facilitating consistent detection. The estimated positions of the six individuals, as determined by the proposed method, are compared to their ground-truth positions. Figure 7(b) displays the reconstructed 3D map with object information. As demonstrated, the indoor environment is accurately reconstructed, with all six individuals correctly detected and clustered using the proposed method.

(a) (b)

Figure 7: (a) Overview of the robot simulation platform; (b) Reconstructed map with object information.

Table 2 presents the performance metrics of the object coordinate estimation. The results suggest that the estimated positions align well with the ground truth. Specifically, the error in the x-direction ranges from 0.03 m to 0.47 m, with an average of 0.24 m. The error in the y-direction spans from 0.01 m to 0.37 m, with an average of 0.28 m. These promising results highlight the potential of the proposed method for object detection and mapping in indoor environments, demonstrating its accuracy and practical applicability.

Table 2: Comparison of estimated and ground-truth object positions.

Object id	Estimation	Ground truth	x error (m)	y error (m)
1	(5.45, 4.02)	(5.5, 4)	0.05	0.02
2	(6.08, -2.72)	(6, -3)	0.08	0.28
3	(-0.03, -4.73)	(0, -5)	0.03	0.27
4	(-4.33, -5.63)	(-4.5, -6)	0.17	0.37
5	(-5.60, 0.66)	(-6, 0.5)	0.4	0.16
6	(-3.53, 4.49)	(-4, 4.5)	0.47	0.01

The performance of the proposed method is further evaluated using real video data collected in a lounge room on the UTK campus. The room is densely populated with over 20 chairs, multiple desks, and assorted furniture, posing a considerable challenge for the proposed method. Figure 8 displays the reconstructed 3D map incorporating object information. The results suggest that the majority of chairs are successfully detected with precise positioning. The small door handle on the right side is accurately identified and localized. It is important to note that, on the left side, some chairs are erroneously detected as sofas due to their similarities, leading to inconsistent predictions from different camera perspectives.

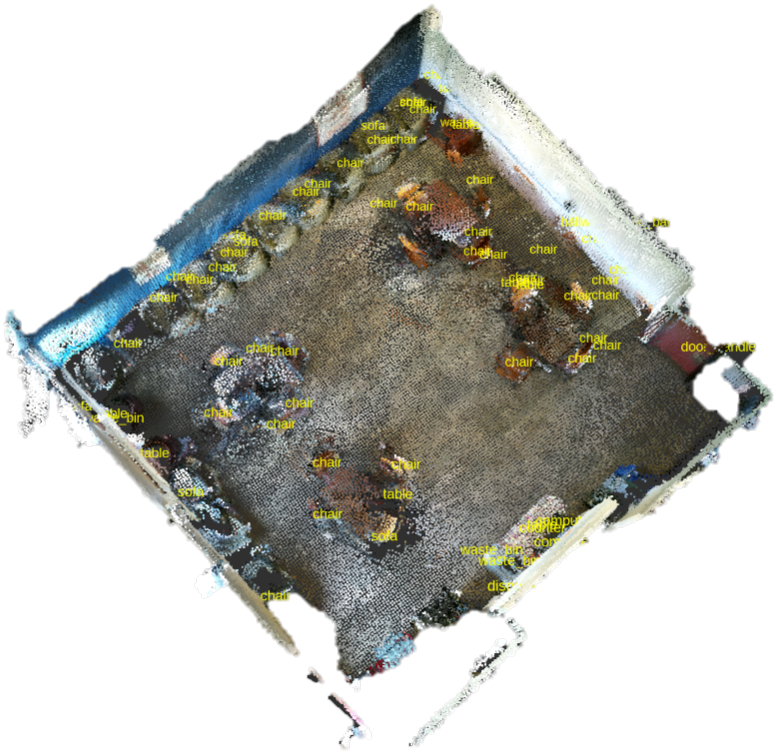


Figure 8: Reconstructed 3D map with object information in a building room.

5 CONCLUSIONS

This study presents a novel approach for indoor object detection and mapping, underpinned by two computational innovations and exhibiting high performance, as validated through real-world data and scenarios. The developed method outperforms existing solutions by accurately detecting and classifying 56 categories of indoor objects in healthcare facilities in real time. This accomplishment was realized through the creation of an unparalleled dataset for robust performance and the incorporation of a new attention mechanism within the deep learning method for detection and classification. The proposed deep learning network achieved an AP_{50} of 67.6% and an mAP of 46.7% on the validation dataset. Furthermore, the proposed method is lightweight, attaining real-time detection with an FPS of 67. Consequently, the AI method can be implemented in an embedded system for real-time detection. A novel mechanism was devised to estimate the 3D coordinates of detected objects and project them onto 3D maps. A robot simulation platform was constructed to assess the performance of the 3D coordinate estimation, yielding an average error of 0.24 m and 0.28 m in the x and y directions, respectively. The methods and workflow were also validated in a real indoor environment on campus, demonstrating their applicability in real-world applications.

Despite the promising results obtained in this study, several limitations should be acknowledged and addressed in future research. First, the dataset used for training and validating the deep learning model was collected from a specific healthcare facility, potentially limiting the generalizability of the method to other settings with different object categories and environmental conditions; expanding the dataset to include diverse healthcare facilities would enhance the generalizability and robustness of the proposed method. Second, the performance of the 3D coordinate estimation mechanism may be influenced by factors such as sensor noise, calibration errors, and varying lighting conditions, which were not extensively explored in this research; future studies could investigate these factors to improve accuracy and reliability in real-world applications. Third, the study focused on static objects, not considering the impact of dynamic objects, such as moving personnel, on the detection and localization process; incorporating techniques for handling dynamic objects, such as real-time tracking and prediction, would enable the system to adapt to rapidly changing environments. Finally, integrating the proposed method into an end-to-end robotic system for disinfection and navigation would allow for a comprehensive evaluation of its effectiveness in real-world scenarios and facilitate the development of more efficient and robust disinfection robots for healthcare facilities.

ACKNOWLEDGMENTS

This research was funded by the US National Science Foundation (NSF) via Grant number 2038967. This research also received support from the Science Alliance at the University of Tennessee Knoxville (UTK) via the Joint Directed Research and Development Program. The authors gratefully acknowledge support from NSF and UTK. Any opinions, findings, recommendations, and conclusions in this paper are those of the authors and do not necessarily reflect the views of NSF, UTK and Kennesaw State University.

REFERENCES

Bradski, G., and A. Kaehler. 2000. "OpenCV". *Dr. Dobb's Journal of Software Tools* 3.

Bashiri, F. S., E. LaRose, P. Peissig, and A. P. Tafti. 2018. "MCIndoor20000: A Fully-Labeled Image Dataset to Advance Indoor Objects Detection". *Data in brief* 17:71–75.

Carling, P. C., M. F. Parry, L. A. Bruno-Murtha, and B. Dick. 2010. "Improving Environmental Hygiene in 27 Intensive Care Units to Decrease Multidrug-Resistant Bacterial Transmission". *Critical Care Medicine* 38(4):1054–1059.

Fuchs, M., and W. Höpken. 2022. "Clustering: Hierarchical, k-Means, DBSCAN". *Applied Data Science in Tourism: Interdisciplinary Approaches, Methodologies, and Applications*, 129–149. Cham: Springer.

Hou, Q., D. Zhou, and J. Feng. 2021. "Coordinate Attention for Efficient Mobile Network Design". In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 13708–13717. Washington, D.C: IEEE Computer Society.

Hu, D., and S. Li. 2022a. "Recognizing Object Surface Materials to Adapt Robotic Disinfection in Infrastructure Facilities". *Computer-Aided Civil and Infrastructure Engineering* 37(12):1521–1546.

Hu, D., J. Chen, and S. Li. 2022b. "Reconstructing Unseen Spaces in Collapsed Structures for Search and Rescue via Deep Learning Based Radargram Inversion". *Automation in Construction* 140:104380.

Hu, D., S. Li, and M. Wang. 2023a. "Object Detection in Hospital Facilities: A Comprehensive Dataset and Performance Evaluation". *Engineering Applications of Artificial Intelligence* 123:106223.

Hu, D., S. Li, J. Du, and J. Cai. 2023b. "Automating Building Damage Reconnaissance to Optimize Drone Mission Planning for Disaster Response". *Journal of Computing in Civil Engineering* 37(3):04023006.

Hu, D., H. Zhong, S. Li, J. Tan, and Q. He. 2020. "Segmenting Areas of Potential Contamination for Adaptive Robotic Disinfection in Built Environments". *Building and Environment* 184:107226.

Huslage, K., W. A. Rutala, E. Sickbert-Bennett, and D. J. Weber. 2010. "A Quantitative Approach to Defining "High-Touch" Surfaces in Hospitals". *Infection Control & Hospital Epidemiology* 31(8):850–853.

Ismail, A., S. A. Ahmad, A. Che Soh, M. K. Hassan, and H. H. Harith. 2020. "MYNursingHome: A Fully-Labelled Image Dataset for Indoor Object Classification". *Data in Brief* 32:106268.

Kejriwal, N., S. Kumar, and T. Shibata. 2016. "High Performance Loop Closure Detection Using Bag of Word Pairs". *Robotics and Autonomous Systems* 77:55–65.

Kinasih, F. M. T. R., C. MacHbub, L. Yulianti, and A. S. Rohman. 2020. "Centroid-Tracking-Aided Robust Object Detection for Hospital Objects". In 2020 *6th International Conference on Interactive Digital Media (ICIDM)*, 1-5. Manhattan, New York City: IEEE.

Labbé, M., and F. Michaud. 2019. "RTAB-Map as an Open-source Lidar and Visual Simultaneous Localization and Mapping Library for Large-scale and Long-term Online Operation". *Journal of Field Robotics* 36(2):416–446.

Liu, S., L. Qi, H. Qin, J. Shi, and J. Jia. 2018. "Path Aggregation Network for Instance Segmentation". In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 8759–8768. Manhattan, New York City: IEEE.

Liu, Y., M. Xu, G. Jiang, X. Tong, J. Yun, Y. Liu, B. Chen, Y. Cao, N. Sun, and Z. Li. 2022. "Target Localization in Local Dense Mapping Using RGBD SLAM and Object Detection". *Concurrency and Computation: Practice and Experience* 34(4):e6655.

Mur-Artal, R., and J. D. Tardós. 2017. "Orb-Slam2: An Open-Source Slam System for Monocular, Stereo, and Rgb-d Cameras". *IEEE Transactions on Robotics* 33(5):1255–1262.

Rosinol, A., A. Violette, M. Abate, N. Hughes, Y. Chang, J. Shi, A. Gupta, and L. Carlone. 2021. "Kimera: From SLAM to Spatial Perception with 3D Dynamic Scene Graphs". *The International Journal of Robotics Research* 40(12–14):1510–1546.

Rublee, E., V. Rabaud, K. Konolige, and G. Bradski. 2011. "ORB: An Efficient Alternative to SIFT or SURF". In *Proceedings of the IEEE International Conference on Computer Vision*, 2564–2571. Manhattan, New York City: IEEE.

Schubert, E., J. Sander, M. Ester, H. P. Kriegel, and X. Xu. 2017. "DBSCAN Revisited, Revisited: Why and How You Should (Still) Use DBSCAN". *ACM Transactions on Database Systems (TODS)* 42(3):1–21.

Vasquez, A., M. Kollmitz, A. Eitel, and W. Burgard. 2017. "Deep Detection of People and Their Mobility Aids for a Hospital Robot". In *2017 European Conference on Mobile Robots (ECMR)*, 1-7. Manhattan, New York City: IEEE.

Wang, C. Y., H. Y. Mark Liao, Y. H. Wu, P. Y. Chen, J. W. Hsieh, and I. H. Yeh. 2019. "CSPNet: A New Backbone That Can Enhance Learning Capability of CNN". In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, 1571–1580. Manhattan, New York City: IEEE.

AUTHOR BIOGRAPHIES

DA HU is an Assistant Professor in the Department of Civil and Environmental Engineering at Kennesaw State University. He holds a Ph.D. degree from the University of Tennessee, Knoxville and master's degree in Civil Engineering from Texas Tech University. His research interests include automation in construction, construction robotics, and disaster response. His e-mail address is dhu3@kennesaw.edu.

MENGJUN WANG Mengjun Wang is a Ph.D. student in the Department of Civil and Environmental Engineering at the University of Tennessee, Knoxville. She graduated from Changsha University of Science & Technology with a bachelor's degree in Traffic Engineering. Her research interests include automation in construction and construction informatics. Her e-mail address is mwang43@vols.utk.edu.

SHUAI LI is an Assistant Professor in the Department of Civil and Environmental Engineering at the University of Tennessee, Knoxville. He graduated from Purdue University with a Ph.D. degree in Civil Engineering, a master's degree in industrial engineering, and a master's degree in economics. He conducts fundamental research in sensing, automation, robotics, and visualization, and applies the techniques in numerous applications, including smart construction, disaster response, and manufacturing. His e-mail address is sli48@utk.edu.