Journal of Engineering, Project, and Production Management
2026, 16(3), 2025-232

Design of an Indoor Visual Positioning System Based on Incremental Learning and Multi-Feature Fusion

Yahui Xie[1] and Zhongping Sun[2]
[1] Associate Professor, School of Art and Design, Fuzhou University of International Studies and Trade, Fuzhou, 350202, China.
[2] Instructor, School of Urban Design, Shanghai Art and Design Academy, Shanghai, 201800 China, E-mail: sunzpingzp@outlook.com (corresponding author).

Project Management
Received October 15, 2025; revised December 8, 2025; accepted December 29, 2025
Available online April 8, 2026

Abstract: Indoor visual positioning in complex indoor spaces requires reliable classification and detection under changing lighting, occlusion, and multi-category interference. To enhance recognition accuracy and real-time performance, a visual orientation system, integrating multi-feature fusion, incremental learning, and a regional proposal detection mechanism, is developed. Deep convolutional, gradient structural features, and texture descriptors are fused to enhance scenario representation, while incremental learning enables category expansion without degrading existing recognition ability. The regional proposal detection module improves target localization and boundary fitting in cluttered indoor layouts. Experimental results on the SUN RGB-D dataset show that the classification model achieves an accuracy of 0.92, a precision of 0.91, a recall of 0.90, an F1-score of 0.91, and a frame rate of 48 frames per second. The detection model achieves an accuracy of 0.94, a precision of 0.92, an Intersection over Union (IoU) score of 0.83, and a processing speed of 38 frames per second. These findings demonstrate that the proposed model achieves a strong balance between recognition performance and operational efficiency, offering practical support for indoor navigation and spatial orientation in dynamic environments.

Keywords: Incremental learning, multi-feature fusion, support vector machine, regional proposal network, visual communication design.

DOI 10.32738/JEPPM-2025-232

1. Introduction

Modern urban spaces are developing rapidly, resulting in large public buildings, transportation hubs, and integrated commercial complexes that feature complex structures and diverse functions. The demand for efficient and intuitive visual communication orientation systems among personnel in such environments is becoming increasingly prominent. Indoor orientation is not only a crucial component of spatial organization but also a vital link to enhancing user experience and safety (Kyle and Downs, 2025). In traditional environments, orientation systems often rely on text, symbols, and manually designed signs to achieve this. However, in scenarios with large spatial scales, dense functional zoning, and frequent personnel flow, a single static symbol often fails to meet the practical needs of crowds for fast positioning, path recognition, and target search (Bouchner et al., 2024). Image recognition and pattern classification technology provides new opportunities for the development of indoor visual communication systems. By automatically recognizing symbols, objects, and scenarios in images, directional signs and environmental information are analyzed in real-time, providing users with more accurate positioning and path planning. However, traditional image classification and detection methods often rely on manual feature extraction, such as directional gradient histograms and Local Binary Patterns (LBP). These methods exhibit limited performance in situations involving lighting changes, complex backgrounds, and target diversity, often resulting in insufficient recognition accuracy and poor generalization capabilities. Some researchers have conducted research on indoor visual communication. Xing et al. (2024) proposed a splicing tampering detection model with a dual-channel enhanced attention mechanism to effectively detect tampered areas in power image splicing, addressing issues such as tampered edge features and poor detection due to a lack of datasets. The research results showed that this method was superior to current models, with improvements in the evaluation index ranging from 1% to 31% and enhanced robustness. Lahoti et al.023) proposed a feature detection method that combined adaptive sampling techniques with a

Convolutional Neural Network (CNN) to detect sparse features of interest. The adaptive sampling explored high-dimensional inputs and utilized regions of interest, while the CNN determined the possibility of feature existence to guide sampling. The research results indicated that this method reduced evaluation time and data processing complexity, while effectively identifying the required features. However, these methods all have low efficiency. Therefore, the research proposes an indoor environment classification model that combines incremental learning and multi-feature fusion. This method combines deep convolutional features, directional gradient features, and LBP features to form a more comprehensive representation of features. Principal Component Analysis (PCA) reduces feature redundancy, and Support Vector Machine (SVM) is used for classification to improve the generalization ability and robustness. This research aims to provide practical and feasible solutions for indoor visual communication orientation systems. Compared with existing hybrid frameworks such as CNN-SVM classification models and RPN-based object detectors, the proposed approach introduces two key advancements. First, the multi-feature fusion mechanism combines deep convolutional features, gradient-based structural cues, and illumination-robust texture descriptors, forming a richer representation space than conventional CNN-only pipelines. Second, the incremental learning strategy enables continual adaptation to new indoor scenario categories without catastrophic forgetting, which is not supported in traditional static CNN-SVM or RPN architectures. These two enhanced functions are integrated into a unified classification detection system, which achieves higher adaptability and stability in constantly changing indoor environments, significantly different from previous hybrid methods.

2. Methods

2.1. Indoor Environment Classification Model based on Multi-Feature Fusion SVM

In modern public buildings and large-scale complexes, indoor visual communication orientation systems have become an important tool for improving space utilization efficiency and user experience. With the increasing complexity of spatial structures, a single text or symbol often fails to satisfy fast positioning and path recognition in unfamiliar environments (Everett et al., 2025). Traditional image classification methods rely on manual feature extraction, which makes it difficult to account for complex factors such as lighting changes, background interference, and target diversity, resulting in insufficient adaptability in dynamic environments. Therefore, an indoor environment classification model based on multi-feature fusion is proposed. As a classification stage in indoor visual communication orientation systems, this model first establishes a feature extraction model based on Visual Geometry Group (VGG) 16, as shown in Fig. 1.

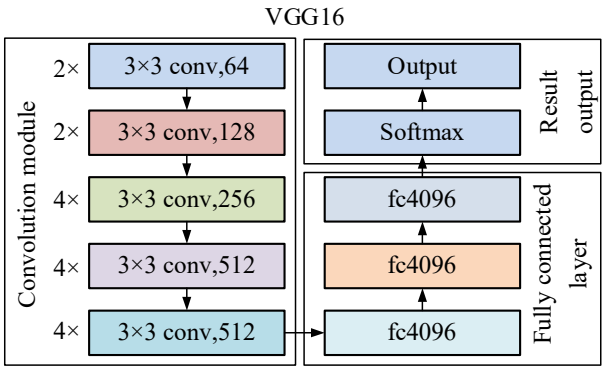


Fig. 1. VGG16 network model diagram

In Fig. 1, VGG16 mainly consists of five sets of convolution modules and three fully connected layers. The convolutional part utilizes 3×3 convolution kernels and extracts feature maps of dimensions 64, 128, 256, and 512 at various levels. To optimize the expressive power of high-level semantic features, both 256-dimensional and 512-dimensional convolution groups achieve deep information extraction through multiple stacking (Kusukawa 2023; Wibisono et al., 2025). The convolution output is flattened and input into three 4,096 dimensional fully connected layers, which are finally mapped to the target category space through a Softmax classifier. The convolution operation is presented in Eq. (1).

$$F_{i,j}^{k}=\sum_{m=1}^{M}\sum_{n=1}^{N}W_{m,n}^{k}\cdot X_{i+m,j+n}+b^{k} \tag{1}$$

In Eq. (1), $F_{i,j}^{k}$ represents the response value of the k-th convolution kernel at position (i,j). $W_{m,n}^{k}$ signifies the convolution kernel weight. $X_{i+m,j+n}$ signifies the pixel value of the input feature map. b^{k} signifies the bias term. The fully connected layer implements linear mapping of high-dimensional features, as presented in Eq. (2).

$$z_j=\sum_{i=1}^{d}w_{ij}\cdot x_i+b_j \tag{2}$$

In Eq. (2), x_i signifies the input feature vector. w_{ij} signifies the connection weight. b_j signifies the bias. z_j

signifies the output of the neuron. In the final classification stage, the Softmax function is used to convert the linear output into a class probability distribution, as expressed in Eq. (3).

$$P(y=c \mid z)=\frac{e^{z_c}}{\sum_{j=1}^{C} e^{z_j}} \tag{3}$$

In Eq. (3), z_c represents the score of category c. C signifies the total categories. $P(y=c \mid z)$ is the probability that the sample belongs to c. Through the hierarchical feature extraction and nonlinear mapping mechanism mentioned above, VGG16 achieves a hierarchical representation of features, progressing from low-level textures to high-level semantics, thereby providing high-precision modeling capabilities for image recognition and classification tasks.

The multi-layer, small convolution kernel structure of VGG16 enables it to capture features from low-level textures to high-level semantics in a step-by-step manner. However, this model also has obvious shortcomings. First, its parameter scale is large, and the training and updating costs are high, making it difficult to retrain frequently in dynamic environments. In addition, VGG16 belongs to the static learning paradigm, and once trained, it cannot efficiently adapt to the addition of new categories or samples. If additional training is directly applied, it often leads to forgetting, meaning the model loses its ability to recognize old tasks while learning new knowledge (Lim et al., 2025; Wang et al., 2024). To address this challenge, the study introduces an incremental learning mechanism. Incremental learning can gradually absorb new data without forgetting old knowledge. To address this, specific strategies were combined, such as freezing low-level convolutions, fine-tuning high-level classifiers, knowledge distillation, and sample replay to achieve continuous optimization and adaptive updating of the model. To stably integrate new scenario categories, the incremental learning mechanism adopts three complementary strategies. First, low-level convolutional layers are frozen to retain previously learned edge, texture, and structural representations, preventing degradation of foundational feature extraction. Second, knowledge distillation is applied to transfer the response characteristics of the original model to the updated classifier, ensuring that newly learned categories do not overwrite past decision boundaries. Third, a sample replay buffer stores a compact set of representative instances from prior classes, allowing the system to rehearse past distributions during incremental updates. The combination of these strategies mitigates catastrophic forgetting, reduces retraining overhead, and preserves long-term classification stability in evolving indoor environments. Unlike conventional CNN-SVM frameworks, which require full retraining when new scenario categories are introduced, the incremental learning mechanism preserves prior decision boundaries while incorporating new samples, thereby eliminating model degradation and reducing retraining costs. To reduce the dimensionality of features, eliminate redundant information, and preserve core features, PCA reduces the dimensionality of the feature matrix, providing more concise feature inputs for classification models. The specific process of PCA dimensionality reduction is shown in Fig. 2.

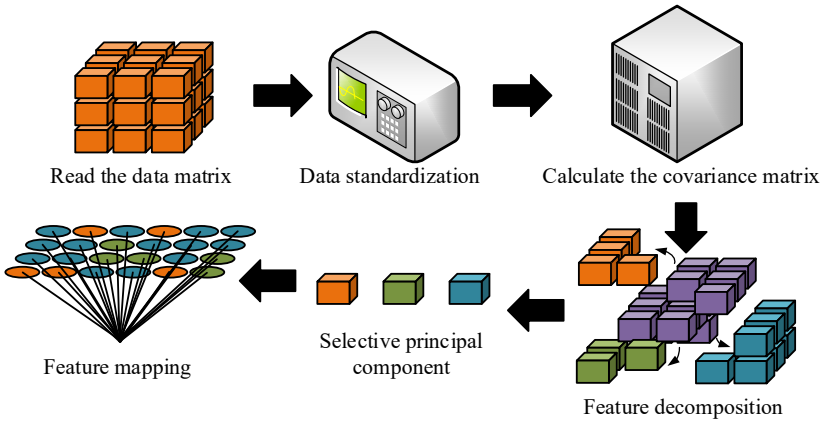


Fig. 2. PCA dimensionality reduction process

In Fig. 2, the PCA data dimensionality reduction process comprises five steps: data standardization, calculation of the covariance matrix, feature decomposition, principal component selection, and feature mapping. First, the original feature matrix is standardized to eliminate the influence caused by different dimensions and numerical ranges. Subsequently, the covariance matrix is calculated to analyze the correlation between each feature. Third, by performing eigenvalue decomposition on the covariance matrix, feature vectors and their corresponding feature values are obtained to characterize the direction and the importance of each principal component (Li et al., 2025). Fourth, based on the cumulative variance contribution rate, a threshold is set to screen key principal components and achieve effective feature selection. Finally, the original data is mapped to a new space composed of the selected principal components to generate a dimensionality-reduced feature matrix. The dimensionality reduction process is shown in Eq. (4).

$$Z=W^T F \tag{4}$$

In Eq. (4), Z represents the reduced-dimensional feature matrix. F represents the feature matrix. W represents the principal component matrix. After completing the dimensionality reduction process, an SVM classification model is constructed based on the reduced feature data to accurately classify different feature dimensions. The optimal hyperplane classification process of SVM is shown in Fig. 3.

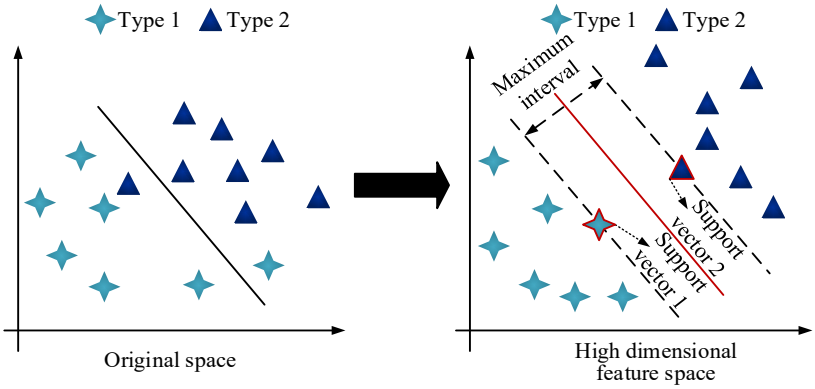


Fig. 3. Schematic diagram of SVM optimal hyperplane

In Fig. 3, the left figure represents the distribution of the original data, with blue triangles and light blue stars representing two types of samples. The two types of data exhibit discrimination in a two-dimensional space, but the boundaries are not clearly defined. The right figure shows the process of introducing the optimal hyperplane into SVM for classification, where the red line represents the classification boundary, which can effectively separate samples of different categories (Nahman-Averbuch et al., 2023; Wang, 2025). The SVM is to find a boundary that maximizes the distance between two types of samples, improving the generalization ability while ensuring correct classification. In this process, the few sample points located on the boundary are called support vectors, which directly determine the position and direction of the hyperplane, while other points far from the boundary have little effect on the result. Fig. 4 presents the model structure.

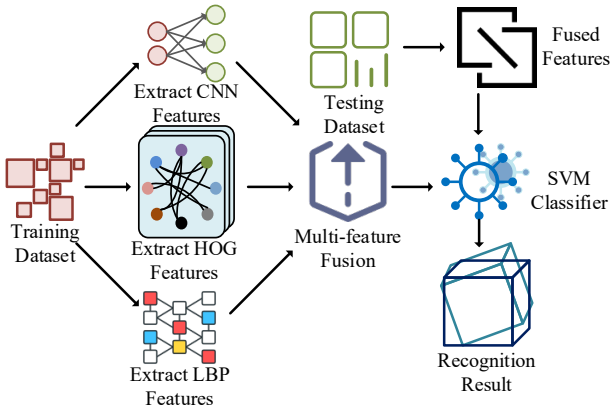


Fig. 4. Indoor visual recognition process based on multi-feature fusion and SVM classification

As shown in Fig. 4, the training data set extracts three types of features: VGG16, directional gradient histogram, and LBP. VGG16 can capture deep semantics and spatial structures, while the Histogram of Oriented Gradients (HOG) highlights the edges and gradient distribution of the target. LBP effectively characterizes local textures and remains stable under lighting changes. The three types of features have their own advantages. After fusion, they form a more comprehensive image expression that includes high-level semantic information while retaining detailed features. The data during the testing phase also undergoes feature extraction and fusion processing to ensure consistency with the feature space during the training phase. The fused features are input into an SVM classification, and its maximum interval discrimination principle is used to achieve accurate partitioning in high-dimensional space, resulting in good generalization ability and stability. The final output recognition result accurately recognizes indoor visual objects.

2.2. Indoor Object Detection Model based on RPN

The classification method based on multi-feature fusion and SVM has achieved fine recognition of indoor objects. However, this method relies on segmented or cropped target regions and cannot directly locate objects in complex scenarios, limiting the overall performance in practical applications. To address this challenge, it is necessary to introduce an efficient object detection mechanism into the indoor visual communication orientation system, which seamlessly integrates the generated candidate regions with subsequent classification. Regional Proposal Network (RPN), as a crucial component of deep

learning object detection, can quickly generate high-quality candidate boxes in an end-to-end framework and share convolutional features with subsequent classifiers, thereby significantly improving detection efficiency and accuracy (Wang, 2025; Liu and Kim, 2025). Therefore, an indoor object detection model is constructed based on RPN. RPN is a deep learning structure used for object detection, whose core task is to automatically generate candidate regions from the entire image that may contain the target. RPN, through end-to-end training, directly predicts the position and target score of candidate boxes based on shared convolutional feature maps, greatly improving detection speed and accuracy. Although RPN-based detectors have been widely applied, their performance in indoor navigation contexts is often constrained by limited by a lack of adaptive scenario expansion. The proposed system differs by coupling RPN detection with a fused feature representation and incremental category expansion, enabling more stable boundary localization under occlusion, clutter, and signage variability. Fig. 5 presents the structure.

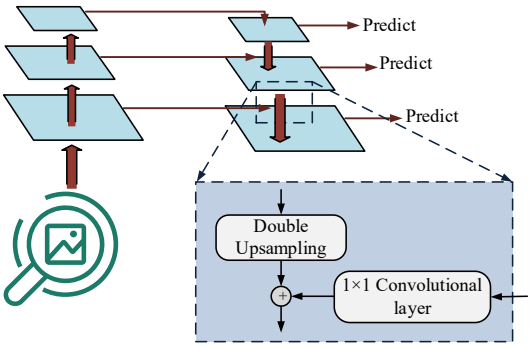


Fig. 5. RPN structure diagram

In Fig. 5, the input image is processed by a CNN to extract multiple layers of feature maps, which contain low-level detail information and high-level semantic information. RPN produces a series of anchor boxes with different scales and aspect ratios at each feature map position, and uses convolution operations to predict whether each anchor box contains the target and the corresponding bounding box regression parameters (Zhang and Su 2025; Ma et al., 2024). In the enlarged module, double up-sampling is used to align the spatial resolution of high-level and low-level features. Then, the convolutional layer is used to compress and map the channel dimension, ultimately achieving feature fusion and predicting candidate boxes. The classification score can be expressed in Eq. (5).

$$p_i = \sigma(w_c^{\cdot} f_i + b_c) \tag{5}$$

In Eq. (5), p_i signifies the probability that the i-th anchor box is the target. σ is the Sigmoid function. w_c signifies the classification weight. f_i signifies the corresponding feature vector. b_c signifies the bias term. The expression for the bounding box regression process is shown in Eq. (6).

$$t_i = w_r^{\cdot} f_i + b_r \tag{6}$$

In Eq. (6), t_i signifies the bounding box regression parameter. w_r and b_r are the regression weight and bias, respectively. The loss function is optimized by combining classification and regression, as expression is shown in Eq. (7) (Heggli et al., 2023; Weber-Lewerenz and Traverso, 2023).

$$L = \frac{1}{N_{cls}} \sum_i L_{cls}(p_i, p_i^*) + \lambda \frac{1}{N_{reg}} \sum_i p_i^* L_{reg}(t_i, t_i^*) \tag{7}$$

In Eq.(7), p_i^* represents the true label. L_{cls} signifies the classification loss. L_{reg} signifies the regression loss. N_{cls} and N_{reg} signify normalization factors. λ is the balance parameter. Through end-to-end feature sharing, up-sampling alignment, and convolutional prediction, RPN can generate high-quality candidate regions, significantly improving detection efficiency and accuracy. The process of the indoor object detection model, based on RPN, is shown in Fig. 6.

In Fig. 6, the input image is first preprocessed, including denoising, normalization, and re-sizing, to ensure image quality and a unified data format. Then, candidate regions are generated through the RPN, automatically generating high-quality candidate boxes on the feature map, which then provides possible target positions for subsequent detection. After obtaining candidate regions, feature extraction is performed on each region, and the extracted features contain both low-level texture and edge information, as well as high-level semantic information, thereby achieving a multi-dimensional representation of objects. These features are input into the classifier to distinguish between target and non-target regions in the candidate area, and the object category is determined. The final output detection result provides the category and location of the target in the indoor scenario.

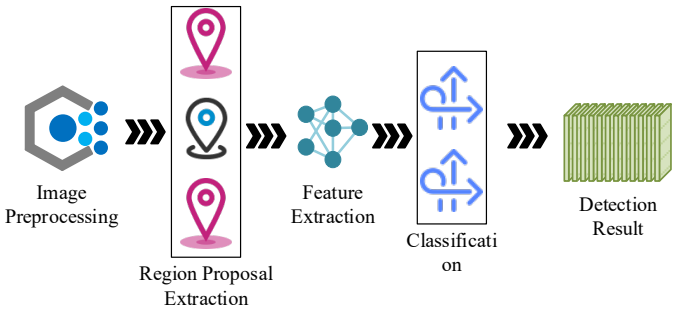


Fig. 6. Process of indoor object detection model based on RPN

In summary, the system designs a comprehensive process, spanning classification to detection, for indoor visual communication orientation tasks. It comprehensively utilizes multi-feature fusion, incremental learning, and SVM classification to achieve environment recognition, and introduces an RPN to improve object detection accuracy and efficiency. In the classification stage, the system integrates deep semantic features of CNN, edge information of HPG, and texture expression of LBP, combined with PCA for feature dimensionality reduction, and finally achieves high-precision classification through SVM. In the detection phase, the system generates high-quality candidate boxes through RPN and shares features with the classifier, achieving fast detection and accurate boundary fitting in complex scenarios.

3. Results

3.1. Performance Analysis of Indoor Environment Classification Model based on Multi-Feature Fusion SVM

The experimental environment configuration used in the study, includes an Intel Core i7-10700K processor, 32GB of memory, and NVIDIA RTX 3080 graphics card. The study takes the SUN RGB-D public dataset, which contains approximately 10,000 RGB-D images collected from various depth sensors, such as Kinect v2, Intel RealSense, and structured light cameras, covering typical indoor environments such as bedrooms, living rooms, kitchens, offices, and shopping malls. This dataset not only provides RGB color images but also provides depth maps and camera pose information. Each image is accompanied by precise object bounding boxes and semantic annotations, covering over 700 types of indoor objects, including tables, chairs, doors, windows, display screens, beds, and more, meeting the multi-task requirements of object detection, semantic segmentation, and scenario recognition. In this experiment, 8,000 images were used for model training, and 2,000 images were used for testing, corresponding to an 80/20 train-test split. The dataset encompasses 45 indoor scenario classes and includes more than 700 labeled object categories. To improve robustness under lighting variations, occlusion, and cluttered environments, data augmentation techniques were adopted, including random rotation (±15°), horizontal flipping, Gaussian noise injection, brightness scaling (±20%), and perspective distortion. These augmentation strategies ensured sufficient intra-class diversity and reduced over-fitting during model learning.

The SVM classifier was configured using a radial basis function kernel. The penalty strength was set to 10, while the kernel scaling factor was set to 0.001. The convergence tolerance was set to 0.0001. A grid search procedure combined with five-fold cross-validation was used to determine the most suitable parameter combination, ensuring balanced precision and recall across indoor scenario categories. For the regional proposal component, anchor sizes were defined at 64, 128, and 256 pixels, with aspect ratios of 1:1, 1:2, and 2:1. The learning rate was initialized at 0.001, with a batch size of 64, momentum set to 0.9, and regularization strength set to 0.0005. The model was trained for 120 epochs with early stopping to prevent overfitting. Non-maximum suppression was applied using an overlap threshold of 0.7 to refine bounding box proposals. The study selects CNN-SVM and HOG-SVM as comparison models, as presented in Fig. 7.

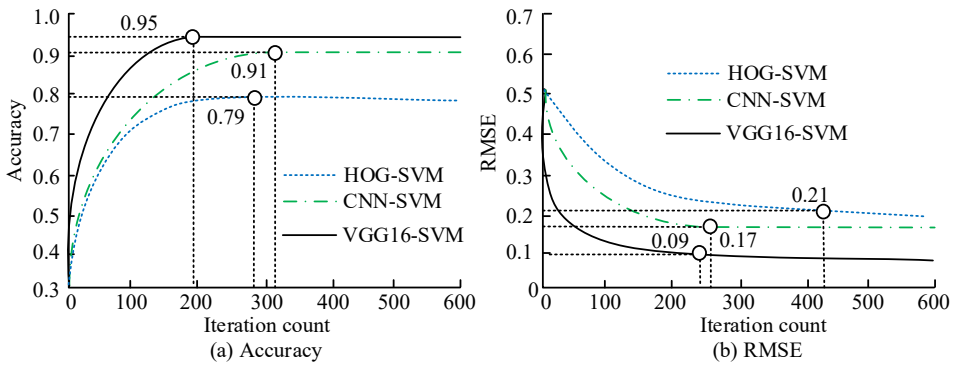


Fig. 7. Changes in accuracy and RMSE

Fig. 7(a) shows the accuracy change for different models as the number of iterations increases, while Fig. 7(b) shows

the Root Mean Square Error (RMSE) changes during the iteration process. According to Fig. 7(a), the VGG16-SVM achieved the highest accuracy of 0.95 after approximately 300 iterations, which was better than that of other models. Deep CNN can extract more discriminative high-level semantic features and achieve strong classification performance when combined with SVM. After about 400 iterations, the CNN-SVM model achieved an accuracy of 0.91, which was inferior to that of VGG16-SVM but significantly better than that of HOG-SVM. This indicates that convolutional features are more robust than traditional gradient features in complex indoor environments. The accuracy of the HOG-SVM model remained around 0.89 throughout the entire process, indicating that relying solely on low-level gradient features is difficult to fully cope with changes in lighting and interference from multiple objects, thus limiting its classification performance. As shown in Fig. 7(b), the VGG16-SVM converged to 0.09 after about 300 iterations, with the lowest error, reflecting its advantages in feature expression and classification boundary determination. The final RMSE of CNN-SVM model was 0.12, slightly higher than that of VGG16-SVM, but still better than that of HOG-SVM. The error of HOG-SVM ultimately stabilized at around 0.17. The proposed model has excellent performance. The performance under different data volumes is analyzed, as presented in Fig. 8.

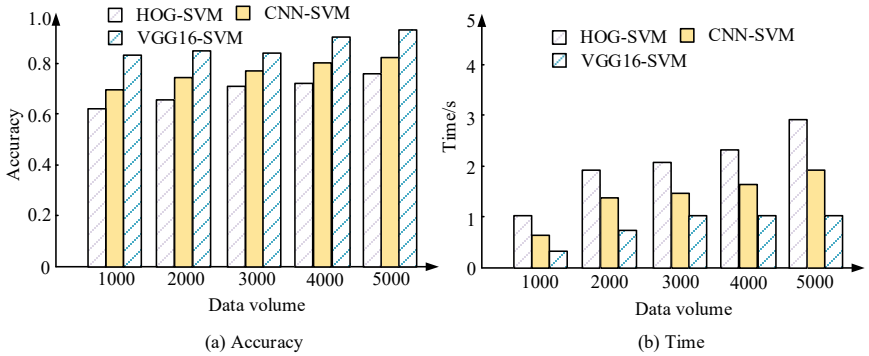


Fig. 8. Comparison of classification accuracy and running time

Fig. 8(a) shows the classification accuracy under different data volumes, while Fig. 8(b) compares the running time. In Fig. 8(a), as the data volume increases, the accuracy of all three models exhibits an upward trend. The VGG16-SVM consistently demonstrated the best performance, achieving an accuracy of 0.82 at a data volume of 1,000 and further improving to 0.92 at 5,000, indicating that its deep convolutional features possess strong expressive and generalization abilities in indoor object classification tasks. The accuracy of CNN-SVM was 0.75 when the data volume was 1,000, gradually improving with the increase of data, reaching 0.87 at 5,000. Although its performance is not as good as VGG16-SVM, it has significantly improved compared with traditional feature models. The accuracy of HOG-SVM was only 0.68 at 1,000, and even when the data size was expanded to 5,000, it was only 0.81. From Fig. 8(b), the computation time of HOG-SVM was consistently the highest, at 1.2 seconds for 1,000 data points, and increased to 3.5 seconds as the data volume grew to 5,000, indicating that manual feature extraction computation is computationally intensive and its scalability is poor with data growth. The CNN-SVM took 0.9 seconds at 1,000 and 1.9 seconds at 5,000, demonstrating relatively high computational efficiency. The VGG16-SVM took 0.8 seconds at 1,000 and only 1.2 seconds even at 5,000. In summary, the VGG16-SVM has low time consumption while ensuring the highest accuracy. The comprehensive performance is analyzed, as presented in Table 1.

Table 1. Comprehensive performance analysis of the model

Model	HOG-SVM	CNN-SVM	VGG16-SVM
Accuracy	0.81	0.87	0.92
Precision	0.79	0.86	0.91
Recall	0.76	0.84	0.90
F1-score	0.77	0.85	0.91
AUC	0.84	0.90	0.95
RMSE	0.17	0.12	0.09
Training time/s	3.5	1.9	1.2
Inference speed (FPS)	25	40	48

The selected metrics provide complementary perspectives on model performance. Accuracy reflects the overall correctness of classification decisions, while precision measures the proportion of correctly identified target classes among predicted positives, indicating reliability in situations where misclassification incurs a cost. Recall evaluates the ability to correctly capture true instances of indoor scenario categories, which is essential in environments where missed detections affect navigation usability. The F1-score balances precision and recall and is particularly meaningful when class

distributions are uneven. The area under the curve measures discrimination capability across thresholds, demonstrating robustness under varying confidence settings. The root mean square error expresses the deviation between predicted and annotated labels, providing a quantitative indicator of prediction consistency. Finally, the number of processed frames per second reflects real-time capability, which is essential for deployment in dynamic indoor spaces with continuous user movement. According to Table 1, the VGG16-SVM performed the best overall in all indicators, followed by CNN-SVM, while HOG-SVM performed relatively weakly. In terms of classification accuracy, the VGG16-SVM achieved an accuracy of 0.92, significantly higher than that of CNN-SVM (0.87) and HOG-SVM (0.81). This indicates that deep convolutional features can effectively extract high-level semantic information in complex indoor environments, improving overall classification performance. For recall and F1-score indicators, the VGG16-SVM achieved 0.90 and 0.91, respectively, both outperforming those of CNN-SVM (0.84 and 0.85) and HOG-SVM (0.76 and 0.77), indicating that the model has more substantial advantages in reducing missed detections and balancing precision and recall. In terms of AUC value, the VGG16-SVM achieved 0.95, surpassing that of CNN-SVM (0.90) and HOG-SVM (0.84), which reflects its advantage in maintaining a stable discriminative ability across different thresholds. The RMSE of VGG16-SVM was only 0.09, which was lower than that of CNN-SVM (0.12) and HOG-SVM (0.17), further proving that its prediction results are closer to the true values. In summary, the VGG16-SVM achieves the best performance in terms of accuracy, robustness, and efficiency.

3.2. Performance of Indoor Object Detection Model based on RPN

To investigate the performance, Fast Region-based CNN (Fast R-CNN) and Single Shot MultiBox Detector (SSD) are compared, as presented in Fig. 9.

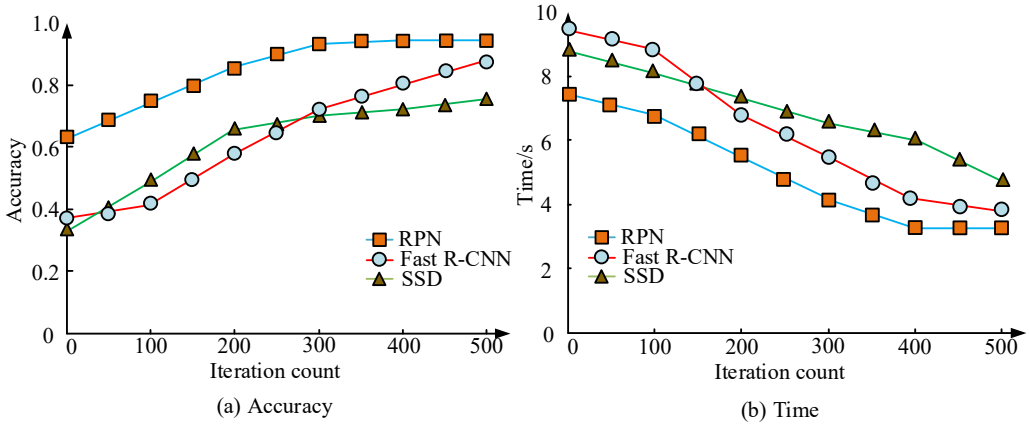
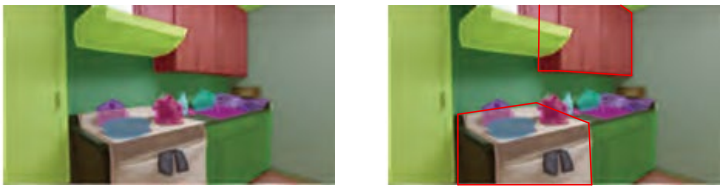


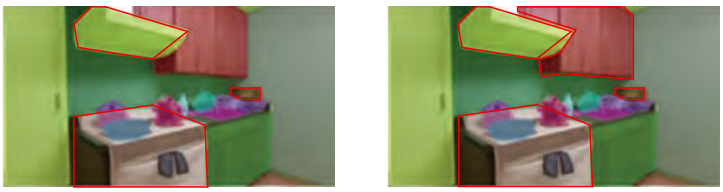
Fig. 9. Accuracy and time consumption of various models

Fig. 9(a) illustrates the changes in accuracy during the iteration process, and Fig. 9(b) shows the time consumption. According to Fig. 9(a), as the iteration increased, the accuracy of all three models gradually improved. The RPN model performed most outstandingly, with an accuracy close to 0.95 at 500 iterations, surpassing that of Fast R-CNN (0.89) and SSD (0.85). This indicates that RPN has higher accuracy and robustness in generating candidate regions and can more effectively extract features of diverse objects in indoor environments. The accuracy of Fast R-CNN remained at an intermediate level throughout the entire process and gradually stabilized with increasing iterations, indicating that its mechanism, which relies on external candidate region generation, limited further improvement in detection performance. In Fig. 9(b), as the iteration increased, the time consumption of all three models gradually decreased. The RPN had the lowest time consumption, dropping to approximately 5 seconds at 500 iterations, which is significantly lower than that of Fast R-CNN (6s) and SSD (7s). It has higher computational efficiency under end-to-end feature sharing and candidate region generation mechanisms. The time consumption of Fast R-CNN remains at an intermediate level, which is related to its reliance on external mechanisms such as selective search, resulting in limited overall computational efficiency. In summary, the RPN not only outperforms Fast R-CNN and SSD in detection accuracy, but also has advantages in computational efficiency, making it more suitable for indoor object detection tasks. The images in the dataset are detected, as presented in Fig. 10.

Fig. 10(a) shows the original image, with typical indoor objects such as cabinets, stoves, and countertops in the scenario. These objects exhibit occlusion and overlap in spatial position, which increases the difficulty of detection. Fig. 10(b) shows the detection results of the SSD model, which can recognize some targets, but the detection box has incomplete boundaries, such as inaccurate positioning of cabinets and boundary deviation. Fig. 10(c) shows the detection results of Fast R-CNN. The detection boxes of this model are more regular compared with SSD, and the boundary recognition of stoves and cabinets is clearer. However, there are still omissions in small area targets. Fig. 10(d) presents the detection results of the RPN. The detection box covers almost all main objects, and the boundaries are more closely aligned with the actual object contours. Especially in cabinet and small object detection, it performs better than SSD and Fast R-CNN. In summary, RPN performs best in terms of accuracy and boundary fitting, making it more suitable for indoor complex object detection tasks. The comprehensive performance of the model is presented in Table 2.



(a) Original image (b) SSD



(c) Fast R-CNN (d) RPN

Fig. 10. Detection performance analysis (Image source: https://rgbd.cs.princeton.edu/)

Table 2. Model detection performance analysis

Model	SSD	Fast R-CNN	RPN
Accuracy	0.85	0.89	0.94
Precision	0.83	0.87	0.92
Recall	0.80	0.84	0.91
F1-score	0.81	0.85	0.91
mAP	0.79	0.84	0.90
IoU	0.70	0.76	0.83
RMSE	0.18	0.14	0.10
Time/s	1.2	2.8	1.9
FPS	45	25	38

According to Table 2, the accuracy of RPN reached 0.94, significantly higher than that of Fast R-CNN (0.89) and SSD (0.85), indicating its stronger overall recognition ability in indoor object detection tasks. The precision and recall of the RPN were 0.92 and 0.91, which were better than those of Fast R-CNN (0.87 and 0.84) and SSD (0.83 and 0.80). This indicates that RPN has a more balanced performance in reducing false positives and false negatives and can better ensure the reliability of detection results. As a comprehensive indicator, the F1-score of RPN was 0.91, which was higher than that of Fast R-CNN (0.85) and SSD (0.81), further proving that it achieved a better balance between accuracy and integrity. The mAP and IoU of RPN were 0.90 and 0.83, respectively, which were better than those of Fast R-CNN (0.84 and 0.76) and SSD (0.79 and 0.70), indicating that its candidate region generation quality was higher, the overlap between the detection box and the real target was greater, and the boundary fitting effect was more accurate. In terms of error, the RMSE of RPN was only 0.10, lower than that of Fast R-CNN (0.14) and SSD (0.18), reflecting that its prediction results are closer to the true values. The RPN not only outperforms SSD and Fast R-CNN in detection accuracy but also considers practicality in terms of speed, making it the model with the best comprehensive performance in indoor object detection tasks. For the detection task, these metrics jointly illustrate the advantages of the proposed model. Higher accuracy, precision, and recall demonstrate stronger reliability in identifying objects such as signage, furniture, and structural markers. A higher F1-score confirms balanced detection stability. The mean average precision and intersection over union indicators reveal tighter alignment between predicted and true bounding regions, especially in scenarios containing occlusion or dense object layouts. A lower error score indicates reduced deviation in bounding positions, while real-time processing capability is evidenced by higher frame throughput, confirming suitability for indoor navigation applications. The performance is analyzed, as presented in Table 3.

According to Table 3, in the standard scenario, the Top-1 accuracy reached 0.93, the Top-5 accuracy increased to 0.97, the specificity and recall exceeded 0.91, the F1-score was 0.91, the average latency was 24ms, and the throughput was 42FPS, fully meeting the real-time requirements. In complex interference scenarios, the Top-1 accuracy dropped to 0.89, the recall was 0.87, and the specificity was 0.88, mainly affected by weak light and occlusion. However, the system still maintained a processing speed of 34FPS and a Top-5 accuracy of 0.95, demonstrating strong robustness. In the incremental learning scenario, the Top-1 accuracy was maintained at 0.90, the recall and F1-score were both 0.89, the calibration error was 0.023, the average latency was 26 milliseconds, and the throughput was 40FPS, indicating that the system maintains

prediction stability while learning new categories. Overall, the system has advantages in accuracy, real-time performance, and adaptability, making it suitable for complex indoor environment applications.

Table 3. Comprehensive performance of the indoor visual communication orientation system in three scenarios

Test Scenario	Standard Scenario	Complex Scenario	Incremental Learning
Top-1 Accuracy	0.93	0.89	0.91
Top-5 Accuracy	0.97	0.95	0.96
Specificity	0.92	0.88	0.91
Recall	0.91	0.87	0.89
F1-score	0.91	0.88	0.89
Calibration Error	0.021	0.031	0.023
Average Latency (ms)	24	32	26
Throughput (FPS)	42	34	40

4. Conclusion

In response to the insufficient classification accuracy and limited real-time performance of indoor visual communication orientation systems in complex environments, a classification method based on incremental learning and multi-feature fusion was proposed. Additionally, an indoor object detection model was constructed using the RPN. The VGG16-SVM achieved an accuracy of 0.92, a precision of 0.91, a recall of 0.90, an F1-score of 0.91, an average AUC of 0.95, a RMSE of 0.09, and an inference speed of 48FPS. In terms of efficiency indicators, the training time of VGG16-SVM was only 1.2 seconds. In the detection task, the RPN model exhibited more prominent performance advantages. The accuracy of RPN was 0.94, the precision was 0.92, the recall was 0.91, the F1-score was 0.91, the average accuracy reached 0.90, the intersection over union was 0.83, and the lowest RMSE was 0.10. The comprehensive comparison shows that RPN not only outperforms Fast R-CNN and SSD in detection accuracy, but also achieves a good balance in efficiency. The classification model, based on incremental learning and multi-feature fusion, can significantly improve directional symbol recognition ability. The system differs from existing hybrid methods by integrating multi-feature fusion with incremental learning to maintain recognition stability when new scenario categories emerge. At the same time, RPN-based detection contributes enhanced boundary precision and robustness in visually complex indoor layouts. This combined structure enables both adaptive classification and high-accuracy detection, which is not simultaneously achieved in prior methods. However, the research still has shortcomings, such as a high dependence on hardware resources for large-scale real-time deployment. To address these limitations, lightweight network architectures such as MobileNet and EfficientNet can be explored to reduce computational load and enable deployment on embedded or low-power devices. Additionally, domain adaptation strategies and feature alignment techniques can be applied to enhance robustness in scenarios affected by severe occlusions, strong contrasts, or non-uniform illumination. These approaches offer practical pathways for enhancing scalability and environmental adaptability in future system development. Future research will focus on lightweight network design, cross-modal data fusion, and adaptive optimization to further enhance the universality and robustness of the system.

Author Contributions

Both authors contributed equally to the conception, methodology, data processing, and writing of the article.

Funding

The research is supported by The Fujian Province Undergraduate Education and Teaching Research Project in 2024, “Reform and Exploration of Cultivating ‘the Craftsman Talents with New Qualities’ Oriented to Digital Preservation and Utilization of Architectural Heritage”, grant number FBJY20240167.

Institutional Review Board Statement

Not applicable.

Declaration of Artificial Intelligence (AI) Tools

The authors used Grammarly solely for language editing and readability improvement. The authors reviewed and verified all content and take full responsibility for the accuracy and integrity of the manuscript.

Reference

Kyle, K. J., and Downs, C. T. (2025). Breeding behavior, visual communication and male combat of Philothamnus occidentalis and Philothamnus natalensis. *The Science of Nature*, 112(2), 22-26. doi: 10.1007/s00114-025-01972-6

Bouchner, P., Novotn, S., Orlick, A., and Topol, L. (2024). Evaluating visual communication interfaces between pedestrians and autonomous vehicles using virtual reality experiments. *Neural Network World*, 34(5), 279-291. doi: 10.14311/NNW.2024.34.015

Xing, J., Tian, X., and Han, Y. (2024). A dual-channel augmented attentive dense-convolutional network for power image splicing tamper detection. *Neural Computing and Applications*, 36(15), 8301-8316. doi: 10.1007/s00521-024-09511-6

Lahoti, G., Ranjan, C., Chen, J., Yan, H., and Zhang, C. (2023). Convolutional neural network-assisted adaptive sampling for sparse feature detection in image and video data. *IEEE Intelligent Systems*, 38(1), 45-57. doi: 10.1109/MIS.2022.3215779

Everett, C. P., Norovich, A. L., Burke, J. E., Whiteway, M. R., Villamayor, P. R., Shih, P. Y., Zhu, Y., Paninski, L., and Bendesky, A. (2025). Coordination and persistence of aggressive visual communication in Siamese fighting fish. Cell Reports, 44(1), 115208.1-115208.21. doi: 10.1016/j.celrep.2024.115208

Kusukawa, S. (2023). Engraving accuracy in early modern England: Visual communication and the Royal Society. *Annals of Science*, 30(6), 601-603. doi: 10.1080/00033790.2023.2240351

Wibisono, A., Song, H. K., and Lee, B. M. (2025). Low-cost visual-based communication for diver-assist AUVs. *IEEE Sensors Journal*, 25(7), 12167-12171. doi: 10.1109/JSEN.2025.3540073

Lim, K. Y., Spencer, E., Bogart, E., and Steel, J. (2025). Speech-language pathologists' views on visual discourse elicitation materials for cognitive communication disorder after TBI: An exploratory study. *Journal of Communication Disorders*, 116(8), 106540.1-106540.11. doi: 10.1016/j.jcomdis.2025.106540

Wang, A. Z., Borland, D., and Gotz, D. (2024). *An empirical study of counterfactual visualization to support visual causal inference. Information Visualization,* 23(2), 1-18. doi: 10.1177/14738716241229437

Li, X., Tse, Y. K., and Bu, X. (2025). Examining corporate social irresponsibility in manufacturing: An eye-tracking study of social media news. *International Journal of Production Economics*, 281(2), 141-149. doi: 10.1016/j.ijpe.2025.109539

Nahman-Averbuch, H., Hughes, C., Hoeppli, M.-E., White, K., Peugh, J., Leon, E., King, C. D., and Coghill, R. C. (2023). Communication of pain intensity and unpleasantness through magnitude ratings: Influence of scale type, but not gender of the participant. European Journal of Pain, 27(10), 1161-1176. doi: 10.1002/ejp.2147

Wang, L., Liu, P., and Liang, Y. (2025). Mobile full-duplex wireless light communication. *Chinese Optics Letters*, 23(2), 20603.1-20603.17. doi: 10.3788/COL202523.020603

Wang, H. P. (2025). Adaptive fusion of multi-cultural visual elements using deep learning in cross-cultural visual communication design. Scientific Reports, 15(1), 28431-28433. doi: 10.1038/s41598-025-13386-5

Liu, W., and Kim, H. G. (2025). The visual communication using generative artificial intelligence in the context of new media. Scientific Reports, 15(1), 11577.1-11577.8. doi: 10.1038/s41598-025-96869-9

Zhang, Z., and Su, Y. (2025). Optimizing visual communication in online classrooms using image processing technology. *Traitement du Signal*, 42(1), 373-379. doi: 10.18280/ts.420132

Ma, K., Lee, S., and Chen, H. (2024). Application of visual communication in image enhancement and optimization of human-computer interface. *Mobile Networks and Applications*, 29(5), 1460-1466. doi: 10.1007/s11036-023-02220-9

Heggli, A., Hatchett, B., and Tolby, L. J. M. (2023). Visual communication of probabilistic information to enhance decision support. *Bulletin of the American Meteorological Society*, 104(9), 1533-1551. doi: 10.1175/bams-d-22-0220.1

Weber-Lewerenz, B. C., and Traverso, M. (2023). Navigating applied artificial intelligence in the digital era: How smart buildings and smart cities become the key to sustainability. *Artificial Intelligence and Applications*, 1(4), 230-243. doi: 10.47852/bonviewAIA32021063

Yahui Xie earned a master's degree in Ornamental Plants and Horticulture from Fujian Agriculture and Forestry University in China in 2012. Currently, he works at Fuzhou University of International Studies and Trade in China, serving as an Associate Professor in the Department of Environmental Design and Deputy Dean of the School of Urban Renewal Industry. His research primarily focuses on urban renewal and interior space design.

Zhongping Sun obtained a Doctor of Arts Management degree from Thonburi University in Thailand in 2024. At present, she is teaching at the School of Urban Design of Shanghai Institute of Art & Design. Mainly responsible for teaching and research in the discipline of interior design. She has edited the textbook "Interior Furnishing Design" and the monograph "Multi-dimensional Design of Interior Soft Furnishings." She is also one of the members responsible for formulating the industry standards for soft furnishings design within the China Building Decoration Association. During the more than two decades she has been deeply engaged in her professional field, she has always adhered to the teaching concept of "the coexistence of theory and practice". On the one hand, she focuses on teaching, integrating the latest industry trends and practical experience into her classroom. Over the years, she has not only guided her students to achieve excellent results in various design competitions but also won numerous honors through her own creative practices. Many of her papers have been published in authoritative and core national journals.