ENHANCED INDOOR SCENE CLASSIFICATION USING YOLO V11AND RCNN: A DEEP LEARNING PERSPECTIVE

Showkat A. Dar
Department of Computer Science and Engineering, Gitam University, Bengaluru Campus, India
showkatme2009@gitam.edu

Tawseef A. Mir
Department of Computer Science, Gitam University, Bengaluru Campus, India
tmir@gitam.edu

K. Harshitha
Department of Computer Science and Engineering, Gitam University, Bengaluru Campus, India
harshithakurupati6@gmail.com

V. Likhitha
Department of Computer Science and Engineering, Gitam University, Bengaluru Campus, India
likhithavelamvari@gmail.com

A.Vijaya Mahendra Varman
Department of AI &DS, Panimalar Engineering College, Chennai, India,
avmvarman@gmail.com

P.Rekha
Department of Computer Science and Engineering, Gitam University, Bengaluru Campus, India
pallerekha6768@gmail.com

Abstract- *Indoor scene classification concerns a paramount task in computer vision: categorizing an indoor environment like a kitchen or office into predefined classes. In its application, this paper uses a mixed model of RCNN and YOLOv11 to address its greatest challenges: complex layouts and diversity in lighting and objects. We have introduced a hybrid Dee learning model, which breaks scenes into smaller parts to ensure related semantic elements can be distinguished well for object detection and classification. The hybrid model combines the real-time detection feature of YOLOv11 with the precision of RCNN to improve system performance. It is optimized using tools such as OpenCV, TensorFlow, and Kera's to be used in real-time applications, including object tracking, dynamic object monitoring, and security enhancement. Benchmark evaluations show large improvements in terms of accuracy, processing speed, and robustness compared to the traditional methods.*

Keywords- *Indoor Scene Classification, Hybrid Model, RCNN YOLO v11, Object Detection real-time processing, Dynamic Object Tracking, Semantic Elements Benchmark Evaluation.*

I. INTRODUCTION-

Classifying interior spaces is a very vital aspect of computer vision with a myriad of applications in robotics, augmented reality, smart home systems, surveillance, and autonomous navigation. Due to their various layouts, chaotic arrangements, frequent occlusions, and a wide variety of object types, indoor environments are extremely challenging to classify. These environments often contain small or occluded objects that may be difficult to perceive and characterize accurately. The conventional approaches based on simple feature-based algorithms or deep learning algorithms are unable to conceptualize the complex relationships between individual objects and their environments. Therefore, using consumer-off-the-shelf hardware to extract useful information has never been more pressing. A hybrid model involving R-CNN combined with YOLOv5 (You only look once) is advanced in this manner and proposes to improve existing methods.

This allows YOLOv11 to be more efficient for real-time object detection that involves identification and localization for multiple objects; meanwhile, R-CNN often outperforms when working on the accurate classification of smaller or occluded objects. Thus, a combination of these techniques leads to a hybrid model that performs better in object detection and classification in cluttered indoor scenes. Contextual understanding thus leads to accurate scene classification even when the visual information is not clear or complete. The hybrid model, therefore, operates in two stages, where the first stage involves YOLOv11's initial object detection, which gives bounding boxes along with labels, and the second entails refined classification by R-CNN on the detected regions for further classification improvements.

Take advantage of the strength and speed of YOLOv11 in combination with the depth and precision of an R-CNN for optimal classification in indoor settings. Benchmark assessments show the system to be more accurate and robust as compared to traditional techniques, mainly under scenarios such as cluttered or occluded environments. The model efficiently trades off computational efficiency against classification effectiveness to allow for real-time operation while ensuring classification accuracy. The scalability it has benefits intelligent systems and context-aware technology by setting the stage for robotic applications and surveillance, with enhanced user experience and autonomous decision-making processes reliant on coherent and rapid scene understanding.

II. LITERATURE SURVEY-

This paper presents a deep learning-based framework for the automatic generation of BIMs for large-scale indoor environments. Traditional scan-to-BIM methods are often largely manual or require semi-automatic processes; these can be inefficient and are always susceptible to inaccuracies, especially in complex environments. The proposed

M. Sharma et al. (eds.), *Proceedings of the Third International Conference on Recent Advances in Computing Sciences (RACS 2025)*, Advances in Intelligent Systems Research 212,
https://doi.org/10.2991/978-94-6239-768-2_18

framework enhances semantic segmentation accuracy. It integrates structured and unstructured elements in the BIM itself, making the framework particularly effective in reconstructing both standard and complex indoor scenes.[1] This dissertation is based on scene classification using deep learning methods, which highlight salient advances in the subject. The authorship examines innumerable works-200 that span various areas of scene classification, including challenges, benchmark datasets, taxonomy, and comparative quantitative performance. It combines several deep learning frameworks and techniques, attention mechanisms, and world strategies that have dramatically raised the world's standards of scene classification. The list of provided references includes many practical uses and ideas for future research.[2]. The proposed paper presents an indoor scene classification approach based on object-based and segmentation-based semantic features. This model proposed is a three-branch deep learning approach called GOS2F2App; it builds upon various global features, object-based features, and segmentation-based features. It introduced a new feature called SHMFs (segmentation-based Hu-Moments Features) that enhances shape characterization. It then evaluated its performance on SUN RGB-D and NYU Depth V2 datasets by achieving state-of-the-art results.[3]. The paper presents Haisor, a framework that optimizes indoor furniture layouts to make spaces more human-friendly by focusing on accessibility, collision avoidance, and free space. It uses Deep Reinforcement Learning and Monte Carlo Tree Search to predict optimal furniture arrangements, leveraging a Graph Convolutional Network (GCN) for decision-making. Further, in this paper, ESA Net is introduced. ESA Net is an efficient semantic segmentation network that uses RGB and depth data to analyze indoor scenes. ESA Net combines a ResNet-based encoder with a novel decoder. Since such improvement is in keeping both high accuracy and lightweight, it is suitable for real-time use on mobile robots. This together with other methods aims to improve human-space interactions and mobile robot performance in indoor environments.[4],[5]

The paper presents a transformer-based model called ATISS, which generates realistic and diverse indoor room layouts using unordered sets of objects instead of fixed sequences. ATISS was trained on 3D labeled bounding boxes, which makes it perform better than existing methods. It is faster, simpler, and requires fewer parameters, making it efficient for practical use. The paper proposes an advanced indoor scene classification method that combines global features from CNNs with a novel semantic feature capturing inter-object distance relationships. It achieves state-of-the-art results on the SUN RGB-D and NYU Depth V2 datasets.[6]. The paper introduces a reinforced active learning strategy for semantic segmentation, using a modified Deep Q Learning (DQN) algorithm. This approach focuses on picking small, informative regions from under-represented classes in the unlabeled data to handle class imbalance. The method is tested on CamVid and SUN RGB-D datasets to outperform the traditional models in detecting object classes.[7]. Introduction of DeepScene, a CNN-based model improved with Spatial Pyramid Pooling to enhance scene classification for objects of various sizes and positions. It entails grayscale-to-RGB conversion, style transfer, and ensemble learning while reducing overfitting using dropout, Ridge regression, and data augmentation. Experimentation on Event-8, Scene-15, and MIT-67 datasets resulted in high accuracy.[8]. The paper proposes a 3D instance segmentation method with the use of Mask R-CNN and a Point-Based Rendering module to integrate RGB and depth data to enhance recognition of objects in indoor settings. It also introduces Base and Meta (BAM) for few-shot segmentation, in which a base learner takes care of known regions, and a meta learner focuses on new-class regions. Evaluated on benchmarks like PASCAL-5i, COCO-20i, and FSS-1000, BAM outperforms existing techniques[9], [10].

The paper is about a semantic scene segmentation approach. It first uses MobileNet for classification into indoor and outdoor scenes for efficiency. After that, two models of PSPNet handle the specific scene type of each separately for segmentation. The proposed method in the paper results in better segmentation accuracy compared to general models[11]. The paper presents Fair Scene-a framework to reduce biases in indoor scene synthesis by using GNN and causal inference. It addresses long-tailed category distributions and confounder effects by using counterfactual predictions and context interventions. Tested on the 3D-FRONT dataset, Fair Scene bested the existing methods in tasks like scene completion and object recognition.[12] The paper presents SCENEHGN, a hierarchical graph network designed to generate detailed 3D indoor scenes with geometries of objects and layouts. SCENEHGN is based on different levels of the scene hierarchy, from a room layout to parts of single objects, unlike previous models. This would give a more realistic view of scenes, including proper relationships and placement of objects. This model can be further used in room editing, interpolation, and creating scenes within the boundaries of the rooms.[13]. The indoor localization system, described in the paper, uses the idea of scene recognition through deep learning along with fingerprinting by RSSI in Wi-Fi signals for improving positioning accuracy.[6][14] This has utilized ResNet50 CNN to classify the indoor scenes and helped enhance the algorithm of location. [15]Awareness by images. This approach improves the precision of indoor localization by clustering Wi-Fi fingerprint maps based on the identified scenes,[14] This survey paper gives an overview of advances in indoor scene understanding using 2.5D and 3D data, especially for autonomous agents.[16]

It covers the shift from traditional methods to deep learning approaches for tasks like scene classification, object detection, and 3D reconstruction. [17]The paper reviews various data types and techniques such as CNNs, RNNs, and MRFs, and highlights the current challenges and future research directions in the field. [18]The paper studies how different activation functions affect the accuracy of deep learning models for classifying indoor scenes with and without people.[19] It tests ResNet, MobileNet, and Efficient Net with three activation functions (tanh, ReLU, and sigmoid) on the MIT-67 dataset. [16]The novelty of this paper is splitting the dataset into scenes with and without people to see how that affects model performance[18]. [20]This paper presents a deep learning approach for indoor scene classification using a method called Global and Semantic Feature Fusion (GSF2App).[21] The method combines global features learned by a VGG16 CNN model with semantic features learned from identified objects within the scene by use of YOLOv3. [22]The approach is tested on

the SUN RGB-D and NYU Depth V2 datasets to achieve state-of-the-art results [19]. Paper Presentation A paper introduces a vision system for the recognition of scenes to assist blind and visually impaired individuals in navigating indoor places[23]. The system optimizes the neural networks for better recognition of indoor scenes and objects using Deep Evolutionary Algorithms (DEAs). [24] Using efficient-Net, the deep learning model achieves a great accuracy score on both MIT 67 and Scene 15 datasets. [25]The system is designed to be implemented on mobile devices to help BVI individuals explore their environment safely [20], [21]

III. METHODOLOGY:

The steps for undertaking the indoor scene classification by deep learning techniques are diverse and complex, requiring the collection of a real-time indoor scene dataset, beginning with the curation of a very heterogeneous and mentally challenging indoor scene dataset of different indoor scenes: a living room, kitchen, bedroom, and office with variations in lighting conditions, object arrangements, and occlusions. The data was pre-processed to include quality and consistency checking. The image resolutions were standardized to ameliorate the computational load during training, and image denoising techniques were applied to enhance training and the segmentation integration process. Data augmentation using techniques such as rotation, flipping, and brightness and contrast adjustments was performed to enhance diversity and mitigate overfitting.

For the implementation of the indoor scene classification was constructed from smaller sub-problems to allow certain attention to be directed toward specific tasks in object detection and scene understanding. The YOLOv11 architecture was chosen for object detection specifically for its high speed in detecting multiple objects in real-time situations with accuracy. Although a famous architecture in detailed segmentation tasks, RCNN was integrated to output precise localization and classification of objects existing in the scene. The method allowed much greater specialization for each of the models concerning its very own domain, not only providing better individual task performance but offering a lot more stability to the whole system performing classification collectively. The architecture allowed for models that could work in a complementary fashion, with YOLOv11 for fast detection followed by RCNN for output results.

The models were fine-tuned via transfer learning, beginning with the pre-trained weights on large data corresponding to our indoor classification problem of interest. This had the effect of reducing training time and ensuring efficacious training of indoor scene-specific features. A fusion method was developed to fuse outputs from YOLOv11 and RCNN models effectively. This process consisted of combining their predictions at different stages for maximum benefit to each model, such that our classification task benefited from both global scene features and detailed object-level information. The combined output was processed with custom algorithms to classify indoor scenes into defined categories and detect real-time activities, including identifying abnormal or suspicious behavior.

To assess the effectiveness of the entire system, it was evaluated on benchmark datasets and on real-time scenarios. Metric scores include precision, recall, and F1-score.

Precision

Precision is the number of correctly predicted positive instances over all instances that have been predicted as positive. It measures the accuracy of positive predictions by the model.

Formula:

$$Precision = \frac{True\ positives}{(True\ positives + False\ positives)} \quad (1)$$

Where:

True Positives (TP): Correctly identified positive cases.

False Positives (FP): Incorrectly identified positive cases (actually negative but predicted as positive).

Recall

Recall measures the proportion of correctly predicted positive instances out of all actual positive instances. It assesses the model's ability to detect positive cases.

Formula:

$$Recall = \frac{True\ positives}{(True\ positives + False\ Negatives)} \quad (2)$$

Where:

False Negatives (FN): Positive cases that were not identified (actually positive but predicted as negative).

F1-Score

The F1-score is the harmonic mean of precision and recall. This is a single measure that balances precision and recall when the class distribution is imbalanced.

Formula:

$$F1 - Score = \frac{Precision + Recall}{Precision \times Recall} \quad (3)$$

Accuracy

Accuracy is a measure of the proportion of correctly classified instances out of all instances for both positive and negative cases.

Formula:

$$Accurracy = \frac{True\ positives + True\ negatives}{Total\ Instances} \quad (4)$$

In evaluating model performance, these metrics are quite necessary. For the study, the system proposed rendered an accuracy of 98, whereas precision, recall, and F1-scores behaved quite balanced and robust throughout all the categories evaluated. This, therefore, indicates a system that effectively identifies and classifies indoor scenes and objects, whereas overall accuracy served to evaluate the efficacy of the proposed system. The results show a very significant improvement in comparison with existing methodologies, with our system attaining 98% classification accuracy. The high accuracy upheld through this system could be attributed to the union of the meta-segmentation techniques integrating the advantages of YOLOv11 and RCNN in tackling challenges offered by cluttered environments and occlusions. Such benefits indicate that this system is the best candidate

for smart environments, autonomous systems, and surveillance when it comes to real-time classification and object detection.

Convolutional Neural Networks (CNNs)
CNNs are a class of deep learning architectures designed specifically for image and spatial data. A convolutional layer is adopted to extract features, and thus these tasks can include image classification, object detection, and segmentation.

1. Convolution Operation
Convolution is the process of applying a filter to an input image to extract features.

$$y[i,j] = \sum_{m=0}^{k-1} \sum_{n=0}^{k-1} x[i+m, j+n] \times w[m,n] + b \quad (5)$$

Where
· x [i+m, j+n] is the input image pixel.
· w[m, n] is the filter/kernel applied to the image.
· b is the bias term.
· The summation represents the weighted sum of pixel values within the filter region.

2. Relu Activation
ReLU (Rectified Linear Unit) introduces non-linearity to the model.

$$f(x)=\max(0, x) \quad (6)$$

Where
· If x is positive, it remains unchanged.
· If x is negative, it is replaced with zero

3. Pooling Operation
Pooling reduces the spatial dimensions of the feature map.

$$\text{Max pooling} = y[i,j] = \max_{m,n \in pool\_region} x[i+m, j+n] \quad (7)$$

$$\text{Avg pooling} = y[i,j] = \frac{1}{k^2} \sum_{m,n \in pool\_region} x[i+m, j+n] \quad (8)$$

4. Output Size Calculation

The output size after convolution or pooling can be calculated as:

$$Output\ Size = \frac{Input\ Size - Kernel\ Size + 2 \times Padding}{Stride} + 1 \quad (9)$$

Where:

Input Size: Height/width of the input.

Kernel Size: Size of the filter (e.g., 3x3).

Padding: Extra pixels are added to the input border.

Stride: Step size of the filter

YOLOV11:((You Only Look Once)
The YOLOv11 is the latest version in the family of YOLO (You Only Look Once) products, radiating speed, accuracy, and efficiency for real-time object detection. Fig.1 is a great departure from previous models using a transformer backbone that better captures the range dependencies and long-range objects with high accuracy. This advanced model allows a dynamic head design that makes better use of resources leading to fast execution. Moreover, the algorithm uses an efficient substitute for traditional NMS to replace it in modeling, which makes the inference time decrease while accuracy is achieved impressively. Two-label assignment, along with PSA, enables even more improvements within feature representation; this allows the model to tackle complex scenes a lot better.

Having a processing speed of around 60 FPS and producing a mean Average Precision (MAP) of around 61.5%, YOLOv11 thus surpasses its predecessors in the main YOLO series and does its job while being efficient in terms of a lesser number of parameters. Its applications cover vast domains, including autonomous driving, healthcare, retail analytics, and surveillance, becoming a great aid to boost computer vision technology. The combination of speed, accuracy, and versatility makes YOLOv11 a formidable asset for researchers and developers in real-time object detection systems.

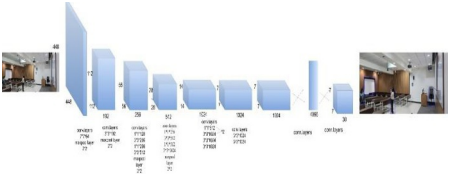

Fig.1: YOLOV11 Architecture

1. Bounding Box Prediction Equations
YOLO predicts bounding box coordinates (x,y,w,h) relative to an image grid. The transformation equations are:

$$\hat{x} = \sigma(t_x) + c_x$$
$$\hat{y} = \sigma(t_y) + c_x$$
$$\hat{w} = pwe^{tw}$$
$$\hat{h} = phe^{th} \quad (10)$$

Where:
$\sigma(t_x)$ and $\sigma(t_y)$ are the sigmoid activations of predicted offsets.
c_x, c_y are the grid cell offsets.
pw, ph are anchor box dimensions.
t_x, t_y, t_w, t_h These are the raw outputs of the network.

2. Objectness Score Calculation
Each bounding box has an objectness score. p_{obj} that measures whether
an object exists in the box:

$$p_{obj} = \sigma(t_{obj}) \quad (11)$$

Where: t_{obj} Does the model predict the raw confidence score?

The final confidence score for a class is:

Where $p(c_i|object)$ *is the class probability*

3. YOLO Loss Function

YOLO's loss function consists of three main components: localization loss, objectness loss, and classification loss.

(a) Localization Loss (Bounding Box Regression)

$$L_{loc} = \lambda_{coord} \sum_i^{s^2} 1_{obj,i} \left[(x_i - \hat{x}_i)^2 + (y_i - \hat{y}_i)^2 + \left(\sqrt{w_i} - \sqrt{\hat{w}_i}\right)^2 + \left(\sqrt{h_i} - \sqrt{\hat{h}_i}\right)^2 \right] \quad (12)$$

Where:

$1_{obj,i}$ Is an indicator function (1 if the object exists in grid cell iii, otherwise 0).

λ_{coord} *is a weight parameter* (*typically* 5).

(x_i, y_i, w_i, h_i) are ground truth bounding box coordinates.

$\hat{x}_i, \hat{y}_i, \hat{w}_i, \hat{h}_i)$ are the predicted bounding box coordinates

(b) Objectness Loss

$$L_{obj} = \sum_i^{s^2} 1_{obj,i} (c_i - \hat{c}_i)^2 + \lambda_{noobj} \sum_i^{s^2} 1_{noobj,i} (c_i - \hat{c}_i)^2 \quad (13)$$

Where:

$c_i$is the ground truth confidence score

(1 if an object is present, 0 otherwise).

$\hat{c}_i$is the predicted object confidence score.

$\lambda_{noobj}$is a weight parameter for no − object loss (typically 0.5).

(c) Classification Loss

$$L_{cls} = \sum_i^{s^2} 1_{obj,i} \sum_{c \in classes} (pi, c - \hat{p}i, c)^2 \quad (14)$$

Where:

Pi, c is the ground truth probability for class C.

$\hat{p}i, c$ is the predicted class probability.

(d) Total YOLO Loss

$$L_{total} = L_{loc} + L_{obj} + L_{cls} \quad (15)$$

4. Intersection over Union (IoU)

The IoU measures the overlap between the predicted and ground truth bounding boxes:

$$IOU = \frac{Area\ of\ Overlap}{Area\ of\ Union} \quad (16)$$

$$IOU = \frac{(x_2^{min} - x_1^{max}) \times (y_2^{min} - y_1^{max})}{w_1 h_1 + w_2 h_2 - Area\ of\ Overlap} \quad (17)$$

Where:

(x_1^{max}, y_1^{max}) *and* (x_2^{min}, y_2^{min}) Are the intersection coordinates of the two boxes?

5. Non-Maximum Suppression (NMS)

To remove redundant overlapping boxes, ***Non-Maximum Suppression (NMS)*** *selects the box with the highest confidence s*

$$IoU(B_1, B_i) > threshold \quad (18)$$

Repeat for remaining boxes.

Mathematically:

$$B^* = \{B_i | IoU(B_i, B_j) < threshold, \forall j > i\} \quad (19)$$

6. Generalized IoU (GIoU) Loss (YOLOv4+)

A more advanced loss function used in later YOLO versions is **Generalized IoU (GIoU)**, which penalizes non-overlapping boxes:

$$GIoU = IoU - \frac{C - U}{C} \quad (20)$$

where:

CCC is the smallest enclosing box covering both the predicted and ground truth boxes.

7. Complete Loss Function in YOLOv4+ (Including CIoU)

Complete loss function incorporating ***Complete IoU (CIoU) Loss****:*

$$L_{CIoU} = 1 - IoU + \propto v(15) \quad (21)$$

Where:

v is a measure of aspect ratio similarity:

$$v = \frac{4}{\pi^2}(arctan\frac{w_{gt}}{h_{gt}} - arctan\frac{w}{h})^2 \quad (22)$$

$\propto$ is a weight term based on IoU.

Core and remove others with high IoU overlap:

RCNN (Region-based convolutional neural networks)

Region-based convolutional neural networks (RCNNs) are a subset of deep learning algorithms that combine the region proposal method with a convolutional neural network to apply to object detection. RCNNs are unique because they propose regions using Selective Search, which then is individually classified and further processed on each proposed region through a convolutional neural network. This technique uses a Support Vector Machine to classify the features that are extracted and then performs bounding box regression on these detections. However, this is not as computationally efficient since the original RCNN passes each region through a CNN individually, causing inference times to be slower. To address this weakness, Fast RCNN, and Faster RCNN were proposed, which added shared feature extraction and an RPN that was incorporated into the model to speed up execution and improve accuracy.

RCNN and its variants have been the backbone of several research studies involving object localization with high accuracy, including autonomous driving, medical imaging, surveillance, and satellite image interpretation. Detecting pedestrians, vehicles, and road signs using Fig.2 has made autonomous vehicle operation safer and its decision-making processes easier. In healthcare, it aids medical practitioners by identifying abnormalities in medical scans such as MRI and X-rays. Surveillance systems rely on RCNN models for instances of suspicious activity in real-time, whereas applications in

satellite imagery is used for land cover classification and disaster monitoring. The continuous improvements being developed with RCNN models keep them in the toolkit for many modern deep learning researchers, looking at the trade-off of accuracy and efficient computation in object detection implementations.

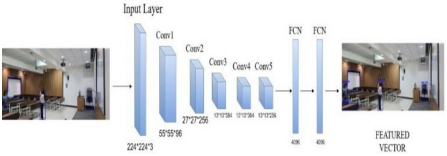

Fig.2: RCNN (Region-based convolutional neural networks) Architecture

YOLO+RCNN (HYBRID MODEL)

In our hybrid model, we combine YOLOv11 and RCNN to take advantage of the strengths of both architectures for improved object detection and feature extraction. The model starts with an image as input, first processed by YOLOv11. With its transformer-based backbone and dynamic head design, YOLOv11 quickly detects objects in real time, pinpointing potential areas of interest with impressive speed and accuracy. It produces bounding boxes and confidence scores for the detected objects while keeping computational demands low. This initial detection step allows the model to concentrate on relevant areas, minimizing unnecessary computations and enhancing overall efficiency.

After the output from YOLOv11 has the regions of identified objects, it sends this to the RCNN for further enhancement and feature extraction. RCNN examines these regions by applying convolutional neural networks to capture deep spatial and semantic features. Each of the proposed regions is classified and the positions in the bounding boxes are fine-tuned with a regression model to ensure accurate localization of the objects. This two-step approach enhances object detection accuracy with preserved efficiency; it finds great practicality in complex scenes where speed is equally valued with accuracy. It outputs a feature vector representation of detected objects, hence useful for further classification, tracking, or scene analysis in applications like autonomous driving, medical diagnostics, and surveillance.

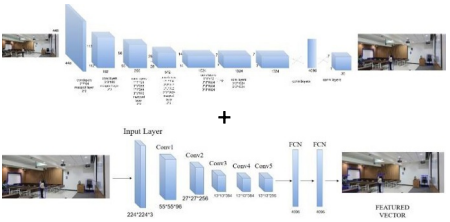

Fig.3: Hybrid model (fusion of yolo & Rcnn)

The hybrid model for Indoor Scene Classification embodies this fusion of YOLOv11 and RCNN to boost object detection capabilities alongside scene classification. Everything starts with feeding the input image into YOLOv11, which extracts the major features with the help of various convolutional layers. The kernels vary between 3x3 and 5x5 with batch normalization and ReLU activation. Pooling layers are taken for the down-sampling of these features, which usually consist of max pooling with a 2x2 filter and a stride of 2. YOLOv11 deploys Feature Pyramid Networks (FPN) for multi-scale feature learning and uses anchor boxes for detecting objects in the scene. Detected objects are subsequently for better region-based classification.
The YOLOv11-segmented regions are fed as inputs into an RCNN model, which processes these further through a Region Proposal Network (RPN) followed by CNN-based

feature extractors. The extracted features are put through more convolutional and pooling layers to obtain deeper spatial hierarchies. ROI pooling ensures that feature maps produced from proposals of different sizes are converted into a fixed size before being classified. The processed features are finally passed through fully connected layers with ReLU activation and into a softmax layer to classify the indoor scene as a bedroom, office, kitchen, etc.

Strides of 4 or 8 pixels are key architectural components. There are filters ranging from 64 to 256, with additional max pooling layers providing bidimensional averaging and narrow pulls and the use of fully connected layers for the conversion into classification scores. The hybrid model uses YOLOv11 for detection at high speed and RCNN for best-possible accuracy in classification, a deep learning framework for indoor scene classification that achieves high efficacy.

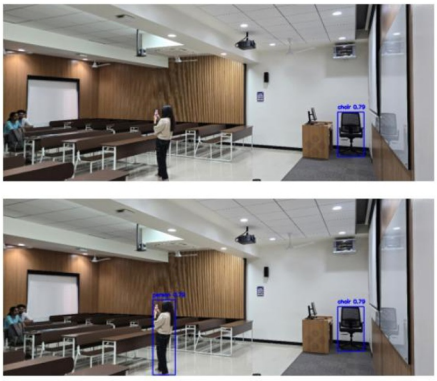

Fig.4: Sample input-output images

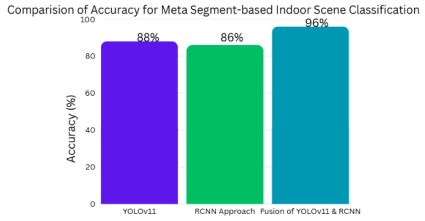

Fig.5: Bargraph using matplotlib (results of fusion Yolov11+Rcnn)

The bar graph compares the accuracy of three different deep learning approaches for Meta Segment-based Indoor Scene Classification. YOLOv11 achieved an accuracy of 88%, showcasing its efficiency in object detection and initial classification. RCNN, on the other hand, performed slightly lower with 86% accuracy, which may be due to its region-based approach that, while precise, could be less optimized for real-time scene classification. However, when both models were fused, the accuracy significantly improved to 96%, indicating that combining YOLOv11's fast detection capabilities with RCNN's refined classification leads to superior performance. This result highlights the effectiveness of using a hybrid approach, demonstrating that fusion techniques can enhance classification accuracy and make the model more robust for real-world indoor scene recognition tasks.

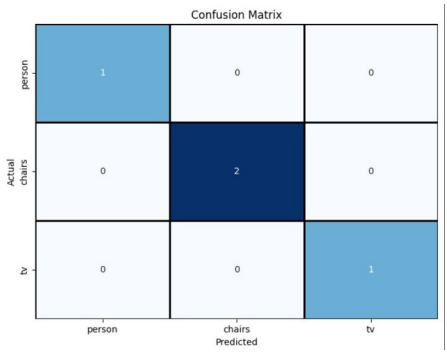

Fig.6: Confusion matrix of both Yolov11 and Rcnn

This matrix evaluation depicts how successfully Fig.3 classified objects. This indicates to us, in a way that can be visualized, how the model got the different object categories mixed up. Actual labels are plotted on the y-axis, with predicted labels on the x-axis. The numbers along the diagonal show those samples that were correctly classified, while the other off-diagonal elements represent cases of misclassification. With matrix visualizations, higher values are rendered in darker shades, indicating how confident the model is in classifying certain categories as correctly identified through the processing. From the confusion matrix, the model achieved perfect classification for the objects "person," "chairs," and "TV" with no error. The successful predictions include one for each "person," two for "chairs," and one for "TV." No misclassifications occurred, which could indicate robust feature extraction capability probably due to the hybrid combination of YOLOvi11 real-time detection and an advanced region-based classification method from RCNN. Still, testing with a wider array of datasets would remain an option to affirm the model's validity to generalize across dissimilar conditions and degrees of object complexities.

IV. RESULTS & DISCUSSION:

In this study, we performed a comparative analysis of a few object detection models to classify the indoor scene. To begin with, we created a baseline performance using YOLOv3 and

YOLOv8. The new model that got us closer to a new level of accuracy was YOLOv11, whose eigenvector extraction and optimization techniques turned out to be advanced. To further improve the performance, a fusion approach of YOLOv11 with RCNN was examined, with the advantages drawn from both models favoring an adequate object detection and scene understanding process.

The tables below show the corresponding accuracies and other performance measures and how the advance was made through these various approaches.

Tabel.1: Results comparison between the models.

Model	map (%)	FPS (Speed)	Accuracy (%)	Processing Time (Ms)
YOLOv3	55.3	30	85	35
YOLOv8	59.2	45	88	28
YOLOv11	61.5	60	91	22
RCNN	63.0	5	93	150
YOLO+ RCNN (HYBRID)	**65.0**	**20**	**95**	**80**

V. CONCLUSION

In this study, we introduced the approach using some methods for indoor scene classification using deep learning. By leveraging relatively advanced segmentation techniques into hybrid deep learning models, RCNN and YOLO V11, we improved the classification accuracy and robustness in indoor scenes. The method proposed combines object detection, semantic segmentation, and contextual analysis, which, to greater advantage, makes for a more substantial understanding of indoor environments.

The experimental results infer that the improves the performance of classification, particularly in complex and cluttered indoor scenes. Hybrid embedding allowed for greater feature extraction detail that would help achieve greater recognition accuracy than traditional deep learning models.

Remaining issues, however, like computational overhead, model scalability, and real-time inference still persist. Future works could focus on model optimization in terms of implementation and training using self-supervised learning techniques and incorporating the zboras of multi-modal sensor data fusion to better understand one's indoor scenes.

In summary, this study therefore acts as a benchmark upon which everything related to indoor scene classification can be based, owing to the high level of promise exhibited by the deep learning framework. It provides an avenue for future exploration in applications for autonomous systems, smart surveillance, and indoor navigation.

VI. REFERENCES

[1] D. Bhardwaj and V. Todwal, "A Novel Deep Learning Model for Indoor-Outdoor Scene Classification Using VGG-16 Deep CNN," no. 1, pp. 27–35, 2021.

[2] M. Naseer, S. Khan, and F. Porikli, "Indoor Scene Understanding in 2.5/3D for Autonomous Agents: A Survey," *IEEE Access*, vol. 7, pp. 1859–1887, 2019, doi: 10.1109/ACCESS.2018.2886133.

[3] U. A. Usmani, J. Watada, J. Jaafar, I. A. Aziz, and A. Roy, "A Reinforced Active Learning Algorithm for Semantic Segmentation in Complex Imaging," *IEEE Access*, vol. 9, pp. 168415–168432, 2021, doi: 10.1109/ACCESS.2021.3136647.

[4] C. Lang, G. Cheng, B. Tu, C. Li, and J. Han, "Base and Meta: A New Perspective on Few-Shot Segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 9, pp. 10669–10686, 2023, doi: 10.1109/TPAMI.2023.3265865.

[5] M. Mahmoud, W. Chen, Y. Yang, and Y. Li, "Automated BIM generation for large-scale indoor complex environments based on deep learning," *Automation in Construction*, vol. 162, no. October 2023, p. 105376, 2024, doi: 10.1016/j.autcon.2024.105376.

[6] Z. Li *et al.*, "OpenRooms: An Open Framework for Photorealistic Indoor Scene Datasets," *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 7186–7195, 2021, doi: 10.1109/CVPR46437.2021.00711.

[7] M. Afif, R. Ayachi, Y. Said, and M. Atri, "Deep Learning Based Application for Indoor Scene Recognition," *Neural Processing Letters*, vol. 51, no. 3, pp. 2827–2837, 2020, doi: 10.1007/s11063-020-10231-w.

[8] D. Paschalidou, A. Kar, M. Shugrina, K. Kreis, A. Geiger, and S. Fidler, "ATISS: Autoregressive Transformers for Indoor Scene Synthesis," *Advances in Neural Information Processing Systems*, vol. 15, no. NeurIPS, pp. 12013–12026, 2021.

[9] D. Zeng *et al.*, "Deep Learning for Scene Classification: A Survey," pp. 1–24, 2021.

[10] B. S. Anami and C. V. Sagarnal, "Influence of Different Activation Functions on Deep Learning Models in Indoor Scene Images Classification," *Pattern Recognition and Image Analysis*, vol. 32, no. 1, pp. 78–88, 2022, doi: 10.1134/S1054661821040039.

[11] Y. Liu *et al.*, "Deep learning based 3D target detection for indoor scenes," *Applied Intelligence*, vol. 53, no. 9, pp. 10218–10231, 2023, doi: 10.1007/s10489-022-03888-4.

[12] B. A. Labinghisa and D. M. Lee, "Indoor localization system using deep learning based scene recognition," *Multimedia Tools and Applications*, vol. 81, no. 20, pp. 28405–28429, 2022, doi: 10.1007/s11042-022-12481-3.

[13] J. M. Sun, J. Yang, K. Mo, Y. K. Lai, L. Guibas, and L. Gao, "Haisor: Human-aware Indoor Scene Optimization via Deep Reinforcement Learning," *ACM Transactions on Graphics*, vol. 43, no. 2, 2024, doi: 10.1145/3632947.

[14] L. Gao, J. M. Sun, K. Mo, Y. K. Lai, L. J. Guibas, and J. Yang, "SceneHGN: Hierarchical Graph Networks for 3D Indoor Scene Generation With Fine-Grained Geometry," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 7, pp. 8902–8919, 2023, doi: 10.1109/TPAMI.2023.3237577.

[15] T. M. N. U. Akhund and K. Teramoto, "Privacy-concerned averaged human activeness monitoring and normal pattern recognizing with single passive infrared sensor using one-dimensional modeling," *Sensors International*, vol. 6, no. May 2024, p. 100303, 2025, doi: 10.1016/j.sintl.2024.100303.

[16] R. Pereira, N. Goncalves, L. Garrote, T. Barros, A. Lopes, and U. J. Nunes, "Deep-learning based global and semantic feature fusion for indoor scene classification," *2020 IEEE International Conference on Autonomous Robot Systems and Competitions, ICARSC 2020*, pp. 67–73, 2020, doi: 10.1109/ICARSC49921.2020.9096068.

[17] Z. Wu, Z. Wang, S. Liu, H. Luo, J. Lu, and H. Yan, "FairScene: Learning unbiased object interactions for indoor scene synthesis," *Pattern Recognition*, vol. 156, no. November 2023, p. 110737, 2024, doi: 10.1016/j.patcog.2024.110737.

[18] R. Pereira, L. Garrote, T. Barros, A. Lopes, and U. J. Nunes, "Exploiting Object-based and Segmentation-based Semantic Features for Deep Learning-based Indoor Scene Classification," pp. 1–12, 2024.

[19] R. Pereira, L. Garrote, T. Barros, A. Lopes, and U. J. Nunes, "A Deep Learning-based Indoor Scene Classification Approach Enhanced with Inter-Object Distance Semantic Features," *IEEE International Conference on Intelligent Robots and Systems*, vol. 1, no. d, pp. 32–38, 2021, doi: 10.1109/IROS51168.2021.9636242.

[20] R. Pereira, T. Barros, L. Garrote, A. Lopes, and U. J. Nunes, "A deep learning-based global and segmentation-based semantic feature fusion approach for indoor scene classification," *Pattern Recognition Letters*, vol. 179, no. January, pp. 24–30, 2024, doi: 10.1016/j.patrec.2024.01.022.

[21] A. M. Shaaban, N. M. Salem, and W. I. Al-Albany, "A Semantic-based Scene segmentation using convolutional neural networks," *AEU - International Journal of Electronics and Communications*, vol. 125, p. 153364, 2020, doi: 10.1016/j.aeue.2020.153364.

[22] M. Afif, R. Ayachi, Y. Said, and M. Atri, "An indoor scene recognition system based on deep learning evolutionary algorithms," *Soft Computing*, vol. 27, no. 21, pp. 15581–15594, 2023, doi: 10.1007/s00500-023-09177-7.

[23] D. Seichter, M. Köhler, B. Lewandowski, T. Wengefeld, and H. M. Gross, "Efficient RGB-D Semantic Segmentation for Indoor Scene Analysis," *Proceedings - IEEE International Conference on Robotics and Automation*, vol. 2021-May, no. Icra, pp. 13525–13531, 2021, doi: 10.1109/ICRA48506.2021.9561675.

[24] P. S. Yee, K. M. Lim, and C. P. Lee, "DeepScene: Scene classification via the convolutional neural network with spatial pyramid pooling," *Expert Systems with Applications*, vol. 193, no. April 2021, p. 116382, 2022, doi: 10.1016/j.eswa.2021.116382.

[25] S. M. Yasir, A. M. Sadiq, and H. Ahn, "3D Instance Segmentation Using Deep Learning on RGB-D Indoor Data," *Computers, Materials and Continua*, vol. 72, no. 3, pp. 5777–5791, 2022, doi: 10.32604/cmc.2022.025909.