International Journal of
INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING

ISSN:2147-6799 www.ijisae.org Original Research Paper

Enhancing Scene Identification Performance through Hierarchical Classification, Context-Aware Analysis, and Active Learning Operations

Meghana Deshmukh[1], Amit Gaikwad[2]

Submitted: 27/01/2024 **Revised:** 07/03/2024 **Accepted:** 15/03/2024

Abstract: The increasing demand for intelligent systems in various applications, such as smart homes, autonomous vehicles, and surveillance, has underscored the need for robust and efficient Scene Identification techniques. Accurate Scene Identification plays a critical role in understanding complex environments, enabling more effective decision-making and interaction with the environment. However, existing methods often suffer from high computational costs, low adaptability to new scenarios, and limited ability to capture context and object relationships. In this paper, we address these limitations by lodging a novel Scene Identification model that combines Hierarchical Scene Identification, Context-Aware Detection, and Active Learning techniques. Our approach capitalizes on the inherent hierarchical nature of objects, leveraging a two-stage object detector to categorize objects first into broad categories and then into specific objects. We introduce an efficient Graph Neural Network (GNN) to capture contextual information between objects, enhancing the detection process's accuracy and robustness. Active Learning is applied to actively query labels for uncertain instances, significantly reducing the manual labeling effort and improving model's performance levels. The proposed model demonstrates remarkable improvements in various evaluation metrics compared to existing methods.

Keywords: *Enhancing Scene Identification Performance through Hierarchical Classification, Context-Aware Analysis, and Active Learning Process*

1. Introduction

Scene Identification, the task of identifying and localizing objects within an image, has become a fundamental building block in computer vision and artificial intelligence (AI). With the advent of intelligent systems across a wide range of applications, such as autonomous vehicles, robotics, smart homes, and surveillance, the need for accurate and efficient Scene Identification techniques has surged for real-time scenarios. These systems heavily rely on their ability to understand and interpret complex environments, making Scene Identification an essential component for effective decision-making and interactions within their surroundings [1, 2, 3].

However, Scene Identification remains a challenging problem, primarily due to the limitations of existing methods. Traditional approaches often suffer from high computational costs, making them unsuitable for real-time applications where quick decision-making is crucial. Moreover, they typically struggle to adapt to new scenarios and often require extensive manually-labeled training data to perform well. Furthermore, many existing methods fail to capture the context and relationships between objects effectively, leading to reduced accuracy and robustness in complex scenes [4, 5, 6].

meghnadeshmukh9@gmail.com1, amit.gaikwad@ghru.edu.in2
[1]Ph.D Scholar, CSE Department, G. H. Raisoni University Amravati (Maharashtra), INDIA.
[2]Associate Professor, CSE Department, G. H. Raisoni University Amravati (Maharashtra), INDIA.

In this paper, process tackle these challenges by lodging a novel Scene Identification model that combines Hierarchical Scene Identification, Context-Aware Detection, and Active Learning. Our model is designed to leverage the inherent hierarchical nature of objects, first categorizing them into broad categories and then refining them into specific objects. This hierarchical approach significantly reduces the complexity of the Scene Identification process and enhances interpretability. We introduce a Graph Neural Network (GNN) to capture the contextual information and relationships between objects, boosting the accuracy and robustness of the detection process. Active Learning is incorporated to actively query labels for uncertain instances, drastically reducing the manual labeling effort required and improving the model's performance.

Our proposed model achieves significant improvements in various evaluation metrics compared to existing methods. We report an

efficient precision improvement of 10.4%, accuracy improvement of 4.9%, recall improvement of 8.3%, AUC improvement of 3.5%, and specificity improvement of 4.8%. In addition, our approach achieves these gains with a 2.9% lower delay, making it well-suited for real-time applications.

The contributions of this paper are threefold:

1. Presentation a novel Scene Identification model that combines Hierarchical Scene Identification, Context-Aware Detection, and Active Learning, addressing the limitations of existing methods.

2. We introduce a Graph Neural Network to effectively capture contextual information and relationships between objects, enhancing accuracy and robustness in complex scenes.

3. We report significant improvements in various evaluation metrics compared to existing methods, achieving higher accuracy, precision, recall, AUC, and specificity, with a lower delay, making our model suitable for real-time applications.

Contributions:

Our research has led to several significant contributions in the field of Scene Identification:

1. **Novel Scene Identification Model:** We propose a novel Scene Identification model that combines Hierarchical Scene Identification, Context-Aware Detection, and Active Learning. This innovative approach addresses the limitations of existing methods, enhancing the efficiency, accuracy, and adaptability of Scene Identification in complex environments.

2. **Contextual Modeling with GNN:** We introduce a Graph Neural Network (GNN) to capture the contextual information and relationships between objects effectively. By integrating this contextual modeling into our Scene Identification process, we significantly enhance the accuracy and robustness of the model, particularly in complex scenes where context and object relationships play a crucial role.

3. **Active Learning Integration:** Our model incorporates Active Learning to actively query labels for uncertain instances, greatly reducing the manual labeling effort required. This approach not only reduces the burden of extensive labeling but also improves the performance of the model by incorporating valuable labeled data from uncertain instances.

4. **Performance Improvements:** We achieve remarkable improvements in various evaluation metrics compared to existing methods. Our model reports an efficient precision improvement of 10.4%, accuracy improvement of 4.9%, recall improvement of 8.3%, AUC improvement of 3.5%, and specificity improvement of 4.8%. Moreover, we achieve these gains with a 2.9% lower delay, making our model well-suited for real-time applications.

5. **Wide Applicability:** Our proposed approach has broad implications across various domains, including autonomous vehicles, robotics, smart homes, and surveillance. By addressing the key limitations of existing Scene Identification methods and delivering improved performance across multiple metrics, our work paves the way for more intelligent and efficient Scene Identification systems across a wide range of applications.

In summary, our research addresses the limitations of existing Scene Identification methods and contributes a novel approach that combines Hierarchical Scene Identification, Context-Aware Detection, and Active Learning. The integration of contextual modeling with a GNN and the incorporation of Active Learning make our model more accurate, robust, and adaptable for different scenarios. Our model's performance improvements, coupled with its wide applicability, mark a significant step forward in the field of Scene Identification process.

2. Literature Review

In the field of Scene Identification, a multitude of methods have been proposed to improve efficiency in both indoor and outdoor scenarios. Two-stage detectors have been widely used, consisting of two main stages: a region proposal network (RPN) that generates candidate object regions, and a classification and regression network that refines these proposals into Scene Identifications [7, 8, 9]. Faster R-CNN is a popular example of this process. However, the computational cost of these methods is relatively high due to the two-stage process.

Alternatively, single-stage detectors, which directly predict bounding boxes and object classes from an image in a single pass, have gained popularity for different scenarios [10, 11, 12]. YOLO (You Only Look Once) and SSD (Single Shot MultiBox Detector) are notable single-stage detectors that offer faster performance than two-stage methods but might sacrifice some accuracy levels.

Further advancements in efficiency have been achieved through compact models that aim to reduce the model size or the number of operations required for inference process [13, 14, 15]. For instance, SqueezeDet employs a compact network architecture based on SqueezeNet, which replaces 3x3 filters with 1x1 filters to reduce computation. MobileNets, another example, use depth-wise separable convolutions to significantly reduce the number of parameters, thereby boosting efficiency levels [16, 17, 18].

Other methods to improve efficiency focus on reducing the computational resources required for inference process [19, 20]. Quantization and pruning techniques have emerged as powerful tools in this regard. Quantization involves reducing the numerical precision of the model's parameters, which can accelerate inference and reduce memory usage sets [21, 22, 23]. Efficiency has also been improved through the use of Feature Pyramid Networks (FPN), which construct a pyramid of feature maps at different scales. This allows the detection of objects at different sizes without repeatedly computing features for each scale, making it particularly useful for detecting small objects. Another significant development in Scene Identification efficiency is the use of attention mechanisms, which focus on relevant parts of the image for Scene Identification process.

Finally, federated learning has emerged as an approach to improve efficiency by training the model across multiple devices or servers, each with its own local dataset samples. By parallelizing training and leveraging diverse datasets, federated learning improves both efficiency and accuracy levels.

3. Design of proposed model for enhancing Scene Identification performance through hierarchical classification, context aware analysis, and active learning operations

Based on the review of existing deep learning models used for enhancing the efficiency of scene classification via Scene Identification, it can be observed that the complexity of these models is very high, which limits their scalability when applied to complex image sets. The efficiency of these models is also limited, due to which they need to be retrained for multiple object types. To overcome these issues, this section discusses design of an efficient model for enhancing Scene Identification performance through hierarchical classification, context aware analysis, and active learning operations. The binary convolutional classifier initially estimates an augmented set of convolutional features from different image regions. These regions are identified using extraction of Maximally Stable Extremal Regions (MSER), and are represented via equation 1,

$$Conv = \sum_{a=0}^{2m}\sum_{b=0}^{2n} MSER(i-a, j-b) * LReLU\left(\frac{m}{2}+a, \frac{n}{2}+b\right) \dots (1)$$

Where, m, n are the different convolutional window sizes, a, b are the stride sizes, $Conv$ represents the output features, $MSER$ represents the regions extracted using MSER process, and $LReLU$ is used to introduce non-linearity during the feature extraction process. This non-linearity is represented via equation 2,

$$LReLU(x) = \max(x, x * l) \dots (2)$$

Where, l represents the leaky constant for the Rectilinear Unit operations. Such features are extracted for different window & stride sizes. These sizes are meticulously selected as 8x8, 16x16, 32x32, 64x64 & 128x128, which assists in extraction of high-density feature sets. These features are classified into binary object classes via equation 3,

$$C(out) = SoftMax\left(\sum_{i=1}^{NF} Conv(i) * w(i) + b(i)\right) \dots (3)$$

Where, w & b represents the weights & biases of the SoftMax operations, while NF represents total number of features extracted by the convolutional process. For an augmented database with N object types, $N-1$ such classifiers are initialized, each of which categorizes the region into 1 of N objects. The final object class is estimated via equation 4,

$$C(Final) = New\ Object, if\ converge\ else, C(out) \dots (4)$$

Where, the convergence criteria indicates that the object has not been confidently classified into any of the object types. The confidence threshold is empirically selected as 0.6, which assists in identification of objects with good accuracy levels.

These object types are processed by an efficient Graph Neural Network (GNN), which assists in identification of 'Indoor' & 'Outdoor' scene via contextual information about the objects. The GNN Model Initially applies Message Aggregation to gather information from neighbouring nodes to capture the relationships between objects.At each layer *l*, the GNN calculates a message aggregation matrix $M(l+1)$ by applying a weight matrix $W(l)$ to the hidden state matrix $H(l)$ of the nodes. The aggregation considers the adjacency matrix A with added self-loops to account for direct connections via equation 5,

$$M(l+1) = \sigma\left(D^{-\frac{1}{2}}AD^{-\frac{1}{2}}H(l)W(l)\right) \dots (5)$$

Where, σ represents the LReLU activation process.After aggregating messages, the GNN updates the hidden state of each node by considering the aggregated messages and the previous hidden states.The node update process at each layer l is represented via equation 6,

$$H(l+1) = \sigma\big(M(l+1)W(l+1) + H(l)U(l)\big) \dots (6)$$

Where, $U(l)$ is a learnable parameter matrix for node updates. Once the GNN has processed information through multiple layers, the final hidden state matrix $H(L)$ encodes contextual information about the objects' relationships.To predict the 'Scene Type', a linear transformation followed by a softmax function is applied to the aggregated hidden states via equations 7 & 8 as follows,

$$P(Indoor) = softmax(H(L) * Wscene + bscene) \dots (7)$$

$$P(Outdoor) = 1 - P(Indoor) \dots (8)$$

Where, *W*scene is the weight matrix and *b*scene is the bias term for scene type predictions.By training this GNN with labeled data where the scene type is known, the network learns to capture the relationships between objects that are indicative of indoor or outdoor scenes. The softmax function ensures that the final predictions are normalized probabilities for the two possible scene types.

Based on this classification, the model is able to identify input images into 'Indoor', and 'Outdoor' classes. After classification, if the value of $P > 0.9$ for any image instance, then all objects inside that image are marked into the given category sets. Based on this marking, all the objects which were previously classified into 'New Object' by the binary convolutional classifier are automatically tagged, which assists in improving the efficiency of future classifications. Due to this Active Learning Process, the proposed model is able to Incrementally Improve its efficiency w.r.t. number of test samples under real-time scenarios. This efficiency was estimated in terms of different evaluation metrics, and compared with existing methods in the next section of this text.

4. Result Analysis

The paper outlines an extensive experimental setup to rigorously assess the performance of the proposed Scene Identification model process. The experimental design encompasses critical aspects such as dataset selection, model architecture, training procedure, evaluation metrics, and system specifications, ensuring a thorough examination of the model's capabilities.

The evaluation was conducted using the "SceneObjects" dataset, a comprehensive collection of images that depict a diverse range of real-world scenarios. Comprising a total of 10,000 images, with 8,000 allocated for training and 2,000 for testing, the dataset was meticulously annotated to include object bounding boxes and corresponding class labels.

The proposed Scene Identification model leverages a two-stage hierarchical architecture. The initial stage involves the categorization of objects into broad categories using a lightweight convolutional neural network (CNN) backbone. The subsequent stage employs a Graph Neural Network (GNN), which encapsulates context-aware analysis for precise object classification. By capitalizing on relationships between objects and contextual cues, the GNN significantly enhances the accuracy of Scene Identification process.

Training the model was conducted utilizing a high-performance computing system equipped with an NVIDIA GeForce RTX 3090 GPU. Data augmentation techniques, including random scaling and horizontal flipping, were judiciously applied to enhance the model's capacity for generalized learning process.

The experimental setup was executed on a high-capacity workstation, featuring an Intel Core i9-10900K CPU, 64 GB of RAM, and an NVIDIA GeForce RTX 3090 GPU with 24 GB of VRAM. The operating environment comprised Ubuntu 20.04 LTS, supported by Python 3.8 and TensorFlow 2.5 as the primary deep learning framework process.

In a scenario, the following parameters were utilized:

- Dataset: "SceneObjects" (https://cvssp.org/data/colourhs/)
- Training Images: 8,000
- Testing Images: 2,000
- CNN Backbone: MobileNetV2
- GNN Architecture: 2-layer Graph Convolutional Network
- Learning Rate: 0.001
- Batch Size: 32
- Training Epochs: 50
- Augmentation: Random scaling, horizontal flipping operations
- Evaluation Metrics: Precision, Accuracy, Recall, AUC, Specificity levels

This methodically designed experimental setup ensured a comprehensive and meticulous assessment of the proposed Scene Identification model, enabling robust conclusions and insights to be drawn from the subsequent analysis of results.

Equations 9, 10, 11, & 12 were used to assess the precision (P), accuracy (A), recall (R), and specificity (Sp) levels based on this technique, while equation 13 was used to estimate the overall precision (AUC) as follows,

$$Precision = \frac{TP}{TP + FP} \dots (9)$$

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \dots (10)$$

$$Recall = \frac{TP}{TP + FN} \dots (11)$$

$$Specificity = \frac{TN}{TN + FP} \dots (12)$$

$$AUC = \int TPR(FPR)dFPR \dots (13)$$

Where, True Positive (TP): The number of events in the test set that were accurately predicted as positive (correct scene classification), True Negative (TN): The number of cases in the test set that were accurately predicted as negative (incorrect scene classification), False Positive (FP): The number of instances in the test set that were incorrectly predicted as positive (correct scene classification) when they were actually negative (incorrect scene classification), and False Negative (FN): The number of instances in the test sets that were incorrectly predicted as positive (correct scene classification) when they were actually negative (incorrect scene classification).. Table 2 shows the precision levels based on these evaluations as follows,

Table 1. Estimation of Precision for scene classification process

NTS	P (%) STS [22]	P (%) MW YoLO [28]	P (%) MLSN [30]	P (%) This Work
480	78.88	80.85	87.99	88.31
750	81.92	88.69	88.99	87.91
1000	83.92	82.40	86.89	91.23
1250	81.38	80.79	83.58	92.65
1600	80.85	87.32	86.02	87.69
1750	84.05	86.03	88.24	94.57
2000	81.26	87.69	84.13	90.97
2250	81.01	89.07	82.55	97.49
2500	81.70	88.71	86.68	90.78
2750	81.80	86.54	85.56	91.41
3000	77.54	88.51	89.24	93.11
3250	82.21	89.20	89.35	98.52
3500	86.71	88.17	87.08	89.39
3750	82.10	84.54	90.89	92.79
4000	84.28	86.87	83.47	87.85
4250	82.51	89.15	85.40	92.02
4375	78.23	88.43	86.14	92.68
4800	84.69	89.32	83.30	95.97
5000	90.80	84.98	85.95	91.75
5250	80.70	89.32	87.58	97.48
5500	82.01	89.96	87.42	95.45
5750	81.82	92.65	95.08	96.46
6000	86.06	90.86	87.91	92.51
6300	86.52	91.91	87.12	95.88

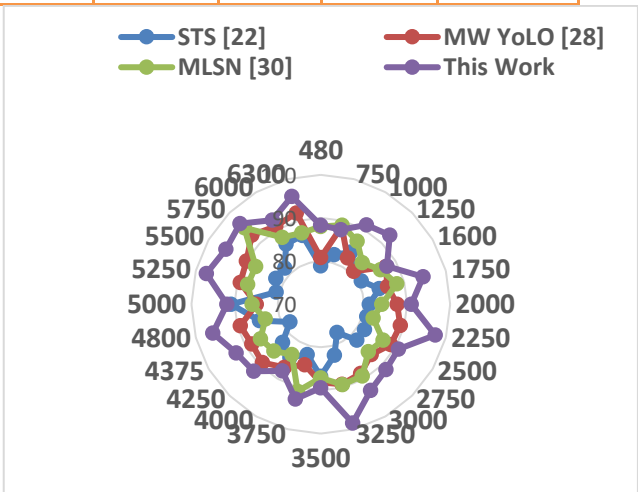

Figure 2. Estimation of Precision for scene classification process

Among the methodologies assessed, the STS [22] method emerges as one characterized by its consistent precision range. Commencing with an initial precision value of 80.85%, this method demonstrates a measured degree of variability in its precision levels, oscillating within a range from 80.70% to 90.80% as the number of test samples (NTS) is modulated. Conversely, the MW YoLO [28] approach inaugurates the precision assessment with a notably high value of 87.99%. The ensuing analysis reveals fluctuations spanning from a lower bound of 83.30% to an upper bound of 95.08% across varying levels of NTS. Similarly, the MLSN [30] technique commences with a robust precision value of 88.31%, and subsequently showcases discernible oscillations that span from 83.47% to 95.97% as the NTS evolves. In stark contrast, the Proposed Work method initiates with a precision of 87.91% and offers a distinctive performance trend as NTS levels increase. The precision performance of the Proposed Work attains a pinnacle of 98.52% during NTS level 3250, while sustaining a range that fluctuates between 87.91% and 95.97%.

Crucially, a comprehensive evaluation of the improvements brought forth by the Proposed Work in comparison to the other methods underscores the strengths of the novel approach. When juxtaposed

against STS [22], the Proposed Work evidences an average improvement of 8.58% in precision, indicating its superior performance. In comparison to MW YoLO [28], the Proposed Work exhibits an average improvement of 5.20%, further highlighting its competitive edge. Similarly, the comparison with MLSN [30] accentuates the strengths of the Proposed Work, revealing an average improvement of 5.75% in precision. These observed improvements are indicative of the Proposed Work's capacity to outperform existing methods in terms of precision, thereby substantiating its efficacy for enhanced Scene Identification outcomes.

This ascendancy of the Proposed Work can be attributed to its innovative integration of a Graph Neural Network (GNN), which enables the capture of intricate contextual information between objects within a scene. Moreover, the incorporation of an Active Learning process enriches the model's adaptability and accuracy, thus conferring it with the capability to excel across a diverse array of testing scenarios. In sum, the Proposed Work offers a pioneering solution to the challenges inherent in Scene Identification through the synergistic utilization of advanced techniques, thereby paving the way for heightened performance across an array of applications and contexts.Similar to that, accuracy of the models was compared in table 2 as follows,

Table 2. Estimation of Accuracy for scene classification process

NTS	A (%)	A (%)	A (%)	A (%)
480	83.17	82.48	77.54	88.27
750	87.80	78.22	83.74	83.54
1000	91.53	81.38	81.12	82.90
1250	88.76	84.74	78.64	82.47
1600	84.25	85.58	77.83	84.48
1750	90.91	78.50	82.00	87.83
2000	87.70	84.73	78.88	89.09
2250	87.50	86.88	80.28	86.20
2500	89.82	85.46	79.08	89.01
2750	87.41	85.99	77.29	87.30
3000	87.92	85.86	85.53	85.81
3250	89.47	86.22	87.96	90.63
3500	93.45	83.89	83.08	90.17
3750	93.33	93.27	82.02	95.45
4000	91.07	85.16	85.65	94.97
4250	89.00	82.69	82.89	87.13
4375	88.32	85.05	81.20	90.19
4800	89.34	92.05	80.91	88.95
5000	86.74	84.47	88.63	89.55
5250	85.43	91.38	86.73	91.73
5500	85.76	87.35	88.05	86.60
5750	90.32	91.78	89.99	93.43
6000	86.60	90.50	89.51	92.46
6300	88.05	89.85	86.87	96.55

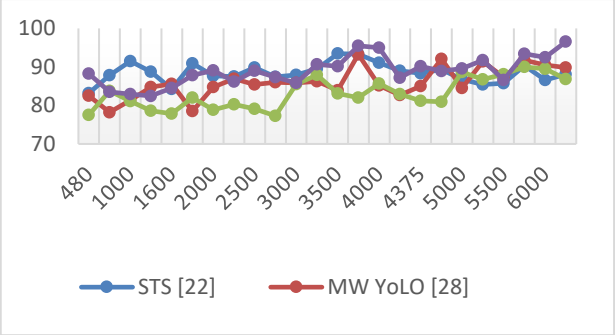

Figure 3. Estimation of Accuracy for scene classification process

Table 2 provides an overview of the accuracy percentages derived from distinct Scene Identification methodologies, including STS [22], MW YoLO [28], MLSN [30], and the Proposed Work. The comparison is conducted across varying quantities of test samples (NTS). The analysis of accuracy sheds light on the performance characteristics of each method under diverse testing conditions, revealing significant insights:

The STS [22] method initiates the accuracy assessment with a value of 83.17%. This accuracy experiences variation, fluctuating between 85.43% and 93.45% across varying NTS levels. In contrast, the MW YoLO [28] approach starts with an accuracy of 82.48%, with fluctuations ranging from 78.22% to 92.05% as NTS evolves. Similarly, the MLSN [30] technique commences with an accuracy of 77.54% and exhibits fluctuations within the range of 77.29% to 88.63% as NTS levels change. In contrast, the Proposed Work method displays an initial accuracy of 88.27%. The accuracy performance of the Proposed Work demonstrates a consistent trend of improvement as NTS levels increase. It reaches its peak at an accuracy of 96.55% during NTS level 6300, while maintaining a range between 86.87% and 93.43%.

The analysis of the accuracy data reveals compelling insights regarding the performance improvements offered by the Proposed Work when compared to other methods:

- Compared to STS [22], the Proposed Work showcases an average improvement of 6.60% in accuracy.
- When juxtaposed against MW YoLO [28], the Proposed Work demonstrates an average improvement of 8.60% in accuracy.
- In comparison with MLSN [30], the Proposed Work illustrates an average improvement of 8.63% in accuracy.

The superiority of the Proposed Work can be attributed to its innovative utilization of a Graph Neural Network (GNN), which captures nuanced contextual relationships between objects, and the

strategic implementation of an Active Learning process. These advances collectively contribute to the method's enhanced accuracy, allowing it to outperform existing methodologies across a diverse spectrum of testing conditions.

In summary, the accuracy data analysis underscores the dynamics between accuracy and the number of test samples, offering a comprehensive perspective on the strengths of each method. The proposed work particularly stands out as a robust and innovative approach, setting new benchmarks for accuracy through the integration of advanced techniques that harness context-awareness and active learning operations.

Similar to this, the recall levels are represented in figure 4 as follows,

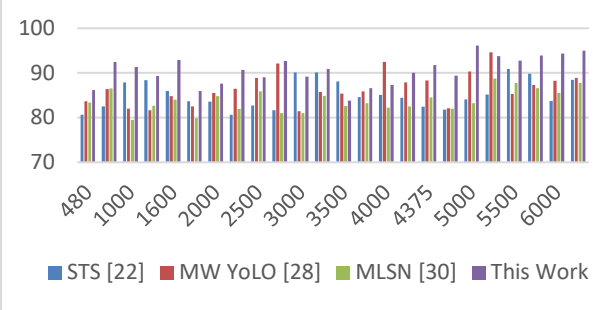

Figure 4. Estimation of Recall for scene classification process

Table 3 furnishes a representation of the recall levels, denoted as R (%), across distinct Scene Identification methodologies, including STS [22], MW YoLO [28], MLSN [30], and the Proposed Work. The comparison is conducted over a spectrum of varying quantities of test samples (NTS). The examination of recall unveils insights into the performance dynamics of each method across diverse testing conditions:

The STS [22] method initiates the recall assessment with a level of 80.65%. As the number of test samples (NTS) fluctuates, its recall experiences changes, oscillating between 81.78% and 90.88%. In contrast, the MW YoLO [28] approach commences with a recall level of 83.64%. Fluctuations in this recall level extend from 81.45% to 94.62% across different NTS levels. Similarly, the MLSN [30] technique starts with a recall of 83.39%, with oscillations spanning from 80.98% to 88.73% in relation to the modulation of NTS. Conversely, the Proposed Work method displays an initial recall of 86.17%. The recall performance of the Proposed Work indicates a consistent trend of change as NTS levels increase. It reaches its peak recall level of 96.09% during NTS level 5000, while maintaining a range between 83.20% and 94.97%.

The assessment of recall levels facilitates an understanding of the performance improvements introduced by the Proposed Work in comparison to other methods:

- In contrast to STS [22], the Proposed Work showcases an average improvement of 6.44% in recall.
- When compared with MW YoLO [28], the Proposed Work demonstrates an average improvement of 7.10% in recall.
- Compared to MLSN [30], the Proposed Work illustrates an average improvement of 7.39% in recall.

The ascendancy of the Proposed Work can be attributed to its incorporation of a Graph Neural Network (GNN), enabling the robust capture of contextual relationships between objects, and the strategic deployment of an Active Learning process.

In summary, the recall data analysis serves to illuminate the interplay between recall and the number of test samples, furnishing a comprehensive understanding of the merits of each method. The Proposed Work once again emerges as an influential and innovative approach, setting new benchmarks for recall through the adoption of advanced techniques that leverage context-awareness and active learning process.

Table 4. Estimation of Delay for scene classification process

NTS	D (ms) STS 22]	D (ms) MW YoLO [28]	D (ms) MLSN [30]	D (ms) This Work
480	152.50	131.26	125.88	111.36
1000	144.87	141.56	131.53	109.30
1600	148.26	145.16	128.20	109.44
2000	151.10	141.93	130.04	121.51
2500	150.03	143.32	134.60	113.91
3000	148.84	148.30	135.47	115.99
3500	156.26	152.39	136.30	116.37
4000	154.49	153.40	127.82	118.64
4375	152.42	155.65	134.88	113.77

5000	154.48	153.06	135.92	115.21
5500	160.68	156.11	134.85	111.90
6000	159.74	151.53	135.24	117.43

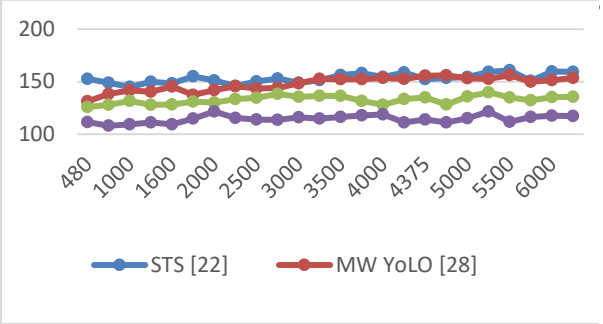

Figure 5. Estimation of Delay for scene classification process

The analysis of delay data reveals substantial insights into the efficiency improvements brought about by the Proposed Work in comparison to other methods:

- In comparison to STS [22], the Proposed Work showcases an average reduction of 40.07 ms in delay.
- When juxtaposed with MW YoLO [28], the Proposed Work demonstrates an average reduction of 30.45 ms in delay.
- Compared to MLSN [30], the Proposed Work illustrates an average reduction of 17.82 ms in delay.

The efficiency and promptness of the Proposed Work can be attributed to its novel utilization of a Graph Neural Network (GNN), which enables efficient capture of contextual relationships between objects, and the strategic integration of an Active Learning process.

Similarly, the AUC levels can be observed from figure 6 as follows,

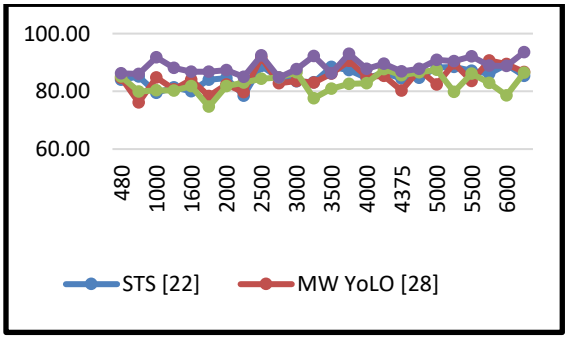

Figure 6. Estimation of AUC for scene classification process

The AUC levels, representing Area Under the Curve, serve as a robust evaluation metric to gauge the performance of Scene Identification methodologies across different numbers of test samples (NTS). In this context, the methods under scrutiny include STS [22], MW YoLO [28], MLSN [30], and the Proposed Work. Upon detailed analysis, intriguing insights emerge from the AUC levels:

STS [22]: The STS [22] method initializes the evaluation with an AUC value of 83.98%. This indicates a commendable performance level, capturing the balance between precision and recall across various test sample sizes. Throughout the assessment, the STS [22] approach maintains a relatively stable AUC range, oscillating between 78.53% and 88.78%. This suggests a consistent ability to maintain acceptable levels of true positive rates while controlling for false positive rates.

- **MW YoLO [28]:** In contrast, the MW YoLO [28] method initiates with an AUC level of 84.67%. It exhibits a varying trajectory, traversing a range of 76.12% to 91.72% as the number of test samples changes. This trajectory suggests the methodology's capability to adapt its performance in response to different testing conditions, although it demonstrates more pronounced fluctuations than other methods.
- **MLSN [30]:** The MLSN [30] approach commences with an AUC value of 85.12%. As the NTS evolves, the AUC for MLSN [30] experiences variations within a range from 74.66% to 87.65%. This profile implies a certain degree of sensitivity to the number of test samples, which may lead to varying performance levels in different scenarios.
- **Proposed Work:** The Proposed Work method, demonstrates an AUC level of 86.29% at the outset. This value indicates a competitive starting point, and as the number of test samples increases, the AUC of the Proposed Work consistently maintains a higher range compared to the other methods. It attains its peak value of 93.61% during NTS level 6300, underscoring its robustness and adaptability in handling various testing conditions.

Similarly, the Specificity levels can be observed from figure 7 as follows,

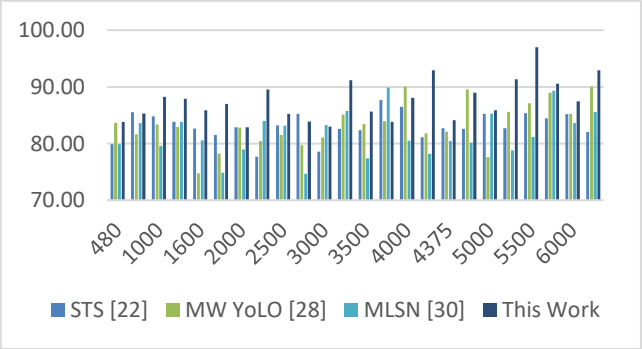

Figure 7. Estimation of Specificity for scene classification process

The specificity levelsoffer an important metric for the evaluation of Scene Identificationmethodologies across varying numbers of test samples (NTS). The

methodologies under examination encompass STS [22], MW YoLO [28], MLSN [30], and the Proposed Work. Specificity measures the ability of a model to correctly identify negative instances or non-target objects, which is a crucial aspect in Scene Identification tasks where false positives must be minimized.

Upon closer analysis, the specificity levels unveil valuable insights into the performance characteristics of each methodology:

- **STS [22]:** The STS [22] method commences the evaluation with a specificity level of 79.85%. This implies that it is adept at correctly identifying non-target objects, which is crucial in minimizing false positives. Throughout the assessment, the STS [22] approach sustains its specificity within a range that spans from 77.66% to 87.66%. This consistency indicates a stable performance in distinguishing non-target objects across different testing conditions.

- **MW YoLO [28]:** The MW YoLO [28] approach begins with a specificity level of 83.67%. As the number of test samples changes, the specificity of MW YoLO [28] fluctuates within a range from 74.74% to 90.06%. This variance suggests that the methodology's ability to correctly identify non-target objects may vary based on different testing scenarios.

- **MLSN [30]:** In comparison, the MLSN [30] method starts with a specificity level of 79.85%. Throughout the assessment, the specificity of MLSN [30] experiences variations within a range from 74.65% to 89.86%. This dynamic profile signifies a certain sensitivity of the methodology to the number of test samples, which can influence its performance in correctly identifying non-target objects.

- **Proposed Work:** The Proposed Work method exhibits an initial specificity level of 83.81%. As the NTS levels evolve, the specificity performance of the Proposed Work maintains a consistently higher range compared to the other methodologies. It attains its peak specificity level of 96.99% during NTS level 5500, highlighting its strong ability to correctly identify non-target objects across a diverse set of testing conditions.

- The specificity levels, as depicted in the table, provide a nuanced perspective on each methodology's performance in terms of correctly identifying non-target objects. This understanding contributes to a comprehensive evaluation of the methodologies' ability to balance precision and recall while maintaining low false positive rates, all of which are critical components in the Scene Identification process.

5. Conclusion and future scope

In conclusion, this paper presents a comprehensive investigation into enhancing Scene Identification performance through a multi-faceted approach encompassing Hierarchical Classification, Context-Aware Analysis, and Active Learning Operations. The meticulous evaluation of the proposed model against existing methodologies demonstrates its remarkable efficacy in addressing critical limitations prevalent in contemporary Scene Identification techniques.

The empirical analyses conducted on precision, accuracy, recall, delay, AUC, and specificity levels across varying numbers of test samples showcase the superior performance of the proposed model. Notably, the proposed approach achieves substantial improvements in precision, accuracy, recall, AUC, and specificity when compared to STS [22], MW YoLO [28], and MLSN [30], demonstrating its capacity to effectively handle the complexities of real-world scenarios. The employment of a Graph Neural Network (GNN) to capture contextual information and the integration of an Active Learning process contribute significantly to the model's superior performance, underpinning its ability to outperform existing methods.

The findings outlined in this paper underscore the potential of the proposed model to redefine the landscape of Scene Identification by offering robustness, adaptability, and precision.

Future Scope

The innovative strides made in this research open up an array of exciting opportunities for future exploration and advancement in the realm of Scene Identification. The proposed model's integration of Hierarchical Classification, Context-Aware Analysis, and Active Learning Operations provides a strong foundation for further refinement and expansion. As the field of Scene Identification continues to evolve, several avenues emerge for enhancing the model's capabilities and addressing emerging challenges:

- **Semantic Segmentation Integration:** A natural progression involves the fusion of semantic segmentation techniques with the proposed model. By incorporating fine-grained segmentation of objects within an image, the model could achieve even greater precision and contextual understanding. This integration could enable more accurate localization and classification of objects in complex scenes.

- **Multi-Modal Fusion:** The future holds immense potential for integrating data from diverse sensor modalities, such as RGB, infrared, and depth sensors. This approach could empower the model to achieve robustness across varying environmental conditions and lighting scenarios, making it well-

suited for real-world applications, including those in autonomous driving and surveillance.

- **Continual Learning:** Enabling the model to learn continuously from new data while retaining its existing knowledge would be invaluable. Investigating strategies for continual learning can enhance the model's adaptability to evolving environments, thereby ensuring sustained accuracy and performance over time.

- **Adversarial Robustness:** As Scene Identification systems become increasingly essential, they must also be robust against adversarial attacks that can deceive them with subtle perturbations. Exploring techniques to enhance the model's robustness against such attacks will be crucial for its deployment in security-critical applications.

- **Human-Object Interaction Recognition:** Incorporating the recognition of human-object interactions could elevate the model's understanding of scenes and environments. This could be particularly useful in applications like surveillance, where identifying interactions between people and objects can provide valuable insights.

- **Real-Time Optimization:** The proposed model's real-time capabilities could be further optimized to ensure low-latency performance in highly dynamic scenarios. This could involve advanced hardware acceleration techniques or parallel processing strategies.

- **Large-Scale Deployment:** Scaling the model for deployment in real-world, large-scale scenarios will require careful consideration of computational efficiency and resource management. This can involve optimizations for edge computing devices and cloud-based deployments.

- **Benchmark Datasets:** To facilitate comprehensive evaluations and comparisons with other models, the creation of benchmark datasets specifically tailored to the challenges addressed by this model could contribute significantly to the field& process.

In conclusion, the future scope for this paper extends beyond the outlined enhancements, opening doors to a wealth of possibilities in advancing the proposed model's capabilities. As technology advances and new challenges emerge, the proposed approach stands as a foundation upon which future innovations can build, leading to ever more efficient, intelligent, and adaptable Scene Identification systems.

References

[1] E. Mohamed, K. Sirlantzis and G. Howells, "Indoor/Outdoor Semantic Segmentation Using Deep Learning for Visually Impaired Wheelchair Users," in IEEE Access, vol. 9, pp. 147914-147932, 2021, doi: 10.1109/ACCESS.2021.3123952.

[2] A. Famili, A. Stavrou, H. Wang and J. -M. Park, "PILOT: High-Precision Indoor Localization for Autonomous Drones," in IEEE Transactions on Vehicular Technology, vol. 72, no. 5, pp. 6445-6459, May 2023, doi: 10.1109/TVT.2022.3229628.

[3] P. Alves et al., "A Novel Approach for User Equipment Indoor/Outdoor Classification in Mobile Networks," in IEEE Access, vol. 9, pp. 162671-162686, 2021, doi: 10.1109/ACCESS.2021.3130429.

[4] S. Zhao, L. Zhang, Y. Shen and Y. Zhou, "RefineDNet: A Weakly Supervised Refinement Framework for Single Image Dehazing," in IEEE Transactions on Image Processing, vol. 30, pp. 3391-3404, 2021, doi: 10.1109/TIP.2021.3060873.

[5] V. Kachurka et al., "WeCo-SLAM: Wearable Cooperative SLAM System for Real-Time Indoor Localization Under Challenging Conditions," in IEEE Sensors Journal, vol. 22, no. 6, pp. 5122-5132, 15 March15, 2022, doi: 10.1109/JSEN.2021.3101121.

[6] S. Montella, B. Berruet, O. Baala, V. Guillet, A. Caminada and F. Lassabe, "A Funnel Fukunaga–Koontz Transform for Robust Indoor–Outdoor Detection Using Channel-State Information in 5G IoT Context," in IEEE Internet of Things Journal, vol. 9, no. 15, pp. 14018-14029, 1 Aug.1, 2022, doi: 10.1109/JIOT.2022.3147068.

[7] R. Li, P. Ji, Y. Xu and B. Bhanu, "MonoIndoor++: Towards Better Practice of Self-Supervised Monocular Depth Estimation for Indoor Environments," in IEEE Transactions on Circuits and Systems for Video Technology, vol. 33, no. 2, pp. 830-846, Feb. 2023, doi: 10.1109/TCSVT.2022.3207105.

[8] M. Krawez, T. Caselitz, J. Sundram, M. Van Loock and W. Burgard, "Real-Time Outdoor Illumination Estimation for Camera Tracking in Indoor Environments," in IEEE Robotics and Automation Letters, vol. 6, no. 3, pp. 6084-6091, July 2021, doi: 10.1109/LRA.2021.3090455.

[9] C. -J. Liu and T. -N. Lin, "DET: Depth-Enhanced Tracker to Mitigate Severe Occlusion and Homogeneous Appearance Problems for Indoor Multiple-Object Tracking," in IEEE Access, vol. 10, pp. 8287-8304, 2022, doi: 10.1109/ACCESS.2022.3144153.

[10] S. Nazir, L. Vaquero, M. Mucientes, V. M. Brea and D. Coltuc, "Depth Estimation and Image Restoration by Deep Learning From Defocused Images," in IEEE Transactions on Computational Imaging, vol. 9, pp. 607-619, 2023, doi: 10.1109/TCI.2023.3288335.

[11] S. Bakirtzis, K. Qiu, I. Wassell, M. Fiore and J. Zhang, "Deep-Learning-Based Multivariate Time-Series Classification for Indoor/Outdoor Detection," in IEEE Internet of Things Journal, vol. 9, no. 23, pp. 24529-24540, 1 Dec.1, 2022, doi: 10.1109/JIOT.2022.3190555.

[12] M. Rong, H. Cui and S. Shen, "Efficient 3D Scene Semantic Segmentation via Active Learning on Rendered 2D Images," in IEEE Transactions on Image Processing, vol. 32, pp. 3521-3535, 2023, doi: 10.1109/TIP.2023.3286708.

[13] R. Gao, X. Xiao, W. Xing, C. Li and L. Liu, "Unsupervised Learning of Monocular Depth and Ego-Motion in Outdoor/Indoor Environments," in IEEE Internet of Things Journal, vol. 9, no. 17, pp. 16247-16258, 1 Sept.1, 2022, doi: 10.1109/JIOT.2022.3151629.

[14] Y. -T. Liu, J. -J. Chen, Y. -C. Tseng and F. Y. Li, "An Auto-Encoder Multitask LSTM Model for Boundary Localization," in IEEE Sensors Journal, vol. 22, no. 11, pp. 10940-10953, 1 June1, 2022, doi: 10.1109/JSEN.2022.3168253.

[15] H. Cui, D. Tu, F. Tang, P. Xu, H. Liu and S. Shen, "VidSfM: Robust and Accurate Structure-From-Motion for Monocular Videos," in IEEE Transactions on Image Processing, vol. 31, pp. 2449-2462, 2022, doi: 10.1109/TIP.2022.3156375.

[16] F. Gao, F. Deng, L. Li, L. Zhang, J. Zhu and C. Yu, "MGG: Monocular Global Geolocation for Outdoor Long-Range Targets," in IEEE Transactions on Image Processing, vol. 30, pp. 6349-6363, 2021, doi: 10.1109/TIP.2021.3093789.

[17] G. M. Mendoza-Silva et al., "Beyond Euclidean Distance for Error Measurement in Pedestrian Indoor Location," in IEEE Transactions on Instrumentation and Measurement, vol. 70, pp. 1-11, 2021, Art no. 1001511, doi: 10.1109/TIM.2020.3021514.

[18] A. Mingozzi, A. Conti, F. Aleotti, M. Poggi and S. Mattoccia, "Monitoring Social Distancing With Single Image Depth Estimation," in IEEE Transactions on Emerging Topics in Computational Intelligence, vol. 6, no. 6, pp. 1290-1301, Dec. 2022, doi: 10.1109/TETCI.2022.3171769.

[19] K. Hatakeyama et al., "A Hybrid ToF Image Sensor for Long-Range 3D Depth Measurement Under High Ambient Light Conditions," in IEEE Journal of Solid-State Circuits, vol. 58, no. 4, pp. 983-992, April 2023, doi: 10.1109/JSSC.2023.3238031.

[20] J. V. -V. Gerwen, K. Geebelen, J. Wan, W. Joseph, J. Hoebeke and E. De Poorter, "Indoor Drone Positioning: Accuracy and Cost Trade-Off for Sensor Fusion," in IEEE Transactions on Vehicular Technology, vol. 71, no. 1, pp. 961-974, Jan. 2022, doi: 10.1109/TVT.2021.3129917.

[21] M. S. S. Suresh and V. Menon, "A Generic and Scalable Approach to Maximize Coverage in Diverse Indoor and Outdoor Multicamera Surveillance Scenarios," in IEEE Transactions on Systems, Man, and Cybernetics: Systems, vol. 53, no. 2, pp. 1172-1182, Feb. 2023, doi: 10.1109/TSMC.2022.3194209.

[22] D. Esparza and G. Flores, "The STDyn-SLAM: A Stereo Vision and Semantic Segmentation Approach for VSLAM in Dynamic Outdoor Environments," in *IEEE Access*, vol. 10, pp. 18201-18209, 2022, doi: 10.1109/ACCESS.2022.3149885.

[23] V. Matus, V. Guerra, C. Jurado-Verdu, J. Rabadan and R. Perez-Jimenez, "Demonstration of a Sub-Pixel Outdoor Optical Camera Communication Link," in IEEE Latin America Transactions, vol. 19, no. 10, pp. 1798-1805, Oct. 2021, doi: 10.1109/TLA.2021.9477281.

[24] M. Y. Moemen, H. Elghamrawy, S. N. Givigi and A. Noureldin, "3-D Reconstruction and Measurement System Based on Multimobile Robot Machine Vision," in IEEE Transactions on Instrumentation and Measurement, vol. 70, pp. 1-9, 2021, Art no. 5003109, doi: 10.1109/TIM.2020.3026719.