AN INDOOR SCENE LOCALIZATION METHOD USING GRAPHICAL SUMMARY OF MULTI-VIEW RGB-D IMAGES

Preeti Meena, Himanshu Kumar, Sandeep Yadav

Indian Institute of Technology Jodhpur, Rajasthan, India

ABSTRACT

The graphical summary of multi-view scene can be readily utilized for tasks such as indoor localization. Existing methods for multi-view indoor localization consider the entire scene for localization purposes by assigning equal importance to all components/objects. In this paper, we propose a novel indoor localization method querying a graphical summary of the scene from the graphical summary of the multi-view RGB-D scenes. Salient objects have been utilized to construct the graphical summaries. The proposed method achieves the best accuracy of 0.90 in scene localization among state-of-the-art methods. Source code is available at https://github.com/preeti-me/MVGSL.

***Index Terms*—** Multi-view RGB-D images, Graphical summary, Scene localization.

1. INTRODUCTION

Indoor location-based services, such as navigation, have led to a demand for accurate indoor scene localization methods. As a result, researchers have been actively exploring and developing various indoor scene localization techniques, including radio-frequency identification (RFID), Bluetooth [1], Wi-Fi [2], and vision-based approaches [3]. In vision-based approaches [4], localization in an indoor scene is performed by matching the visual features. However, only visual features-based matching may result in sub-optimal localization due to variations in visual features arising from illumination, blur, occlusion, etc. Furthermore, visual features alone do not capture the semantic and contextual information of the scene which might be crucial for the scene identification/localization [5].

Alternatively, a summary, a compressed representation of the original scene that accurately conveys its meaningful information, can also be utilized for the localization. The summarized scene not only captures the visual essence of the scene but also can be utilized for the extraction of vital contextual and semantic information of the scene for localization. In this paper, we utilize a graphical summary of the input to identify the scene from the available graphical summary of the multi-view RGB-D scenes in the dataset. Saliency [6] is utilized to generate the representative summary of the scene which in turn is represented as graphical summaries by employing the semantic scene information. The key contributions of this work are: i) A method for graphical summary generation from the multi-view RGB-D images of an indoor scene; ii) A novel graphical summary-based method for indoor scene localization for an RGB-D input. We demonstrate with our analysis that the proposed method obtains the best scene identification accuracy i.e. 0.90 among all the state-of-the-art methods. Also, the proposed method is robust to the noise, blur, occlusions, and different views as per our analysis.

2. RELATED WORKS

Image-based scene localization was introduced by Robertson et al. [3] that uses a database of building facades views to calculate the position of a query image. Afterwards, many image-based scene localization methods have been proposed [7, 8, 9, 10, 11]. Recently, Lu et al. [7] proposed a multi-view localization system that provides the rough location and orientation of the user. Song et al. [12] proposed a method that includes the co-occurring frequency of object-to-object relation (COOR) and sequential representation of object-to-object relation (SOOR) consisting of objects and their spatial relations. Chiou et al. [4] proposed a neural network-based architecture, Graph Location Networks (GLN), based on Graph Convolutional Networks (GCN). Furthermore, the authors use a zero-shot learning approach to reduce labor costs, allowing the system to be employed in large-scale indoor environments. Du et al. [13] proposed a Translate-to-Recognize Network (TRecgNet) that utilized a cross-modal pyramid translation strategy for RGB-D scene identification. Miao et al. [5] proposed an Object-to-Scene (OTS) method, which extracts object features and learns object relations to identify indoor scenes. Zhou et al. [14] implemented an improved object model (IOM) enriched from a Bayesian perspective (BIOM) to find object co-occurrences and pairwise object relations for scene localization. Mosella et al. [15] employed a 2D-3D Geometric Fusion Network that exploits the intrinsic geometric information of the 3D-space to obtain geometric features and improves the fusion with the texture features. More recently, Labinghisa et al. [16] proposed an image-based indoor location awareness algorithm (IILAA) in combination with a clustering algorithm to identify not only the exact spatial location but also the scene cluster of the user. Caglayan et al. [17] im-

plemented a two-stage framework that mapped CNN-based RGB-D features extracted at multiple levels into high-level representations through a fully randomized structure of recursive neural networks. However, it only relied on shape cues and did not consider object semantic features. Girdhar et al. [18] adopted a Transformer architecture for handling multiple modalities in a single encoder. However, the method did not leverage multi-view information. Pereira et al. [19] introduced a two-branch CNN-based Global and Segmentation-based Semantic Feature Fusion Approach (GS^{2}F^{2}App) for scene identification. However, the method did not include the semantic information about the number of objects and their respective categories available in the scene. In contrast, the proposed method utilized the semantic features of salient objects and both intra and inter-view relations between salient objects of multi-view RGB-D images.

3. PROBLEM FORMULATION

A given scene $\mathcal{S}^g$ to be localized in the given database $\mathcal{D}$ consisting of a collection of scenes $\mathcal{S}_i \in \mathcal{B}$ (multi-view RGB-D images) from a building $\mathcal{B}$. Every scene $\mathcal{S}_i$ is represented as a multi-view graph (i.e., graphical summary of multi-view RGB-D images) $\{G_i^{\mathcal{D}}\}_{i=1}^{\mathcal{C}}$ belonging to the $\mathcal{C}$ different scene locations (i.e., rooms) of the given building $\mathcal{B}$. The graph (graphical summary) G^q of a query scene location $\mathcal{S}^g$ is utilized for localization in the building $\mathcal{B}$. The problem has been formulated as given in 1, 2, and 3 below.

Candidate Matching:
$$\hat{i} = \arg\max_{i} \ \mathcal{J}(G^q, G_i^{\mathcal{D}}) \tag{1}$$

Node Alignment:
$$\mathcal{G}_i^c \subseteq \mathcal{G}_i^{\mathcal{D}} \ \text{s.t.} A_{\mathcal{G}_i^c}(k,j) = \begin{cases} 1 & \text{if } j = \arg\max_j \mathbf{d}^{\hat{i}}(f_k, \tilde{f}_j) \\ 0 & \text{Otherwise.} \end{cases} \tag{2}$$

Scene Graph Matching:
$$\hat{i} = \arg\max_{\hat{i}} \ \mathcal{M}(G^q, G_{\hat{i}}^c) \tag{3}$$

Where $\mathcal{J}(.)$ is the Jaccard similarity between the nodes label sets of query graph G^q and graphs in a database $G_i^{\mathcal{D}}$ resulting scene $\hat{i}$ with maximum similarity score. $\mathcal{G}_{\hat{i}}^c$ is matched candidate subgraph with matching $A_{\mathcal{G}_{\hat{i}}^c}$ of nodes of G^q in $\mathcal{G}_{\hat{i}}^c$ such that the similarity metric $\mathbf{d}^{\hat{i}}(.)$ between the feature vectors f_k and $\tilde{f}_j$ of k_{th} and j_{th} nodes of G^q and $\mathcal{G}_{\hat{i}}^c$ respectively. The $\mathcal{M}(.)$ denotes the similarity metric between G^q and weighted subgraph $\mathcal{G}_{\hat{i}}^c$.

4. METHODOLOGY

Figure 1 shows the proposed framework for scene localization comprising of two major stages, namely *Graphical Summary Generation*, and *Scene Localization*. Firstly, we create a dataset containing graphical summaries of multi-view RGB-D scenes which are subsequently utilized for the scene localization as given in the following subsections.

4.1. Graphical summary generation

We combine single-view graphical summaries of the scene to create the multi-view graphical summary of the scene as follows:

A. Single-view graph construction: We first detect the salient objects $\{\mathcal{O}_i\}_{i=1}^n$ with labels $\{L_i\}_{i=1}^n$ in given RGB-D image for view v_i. Then the graphical summary $G^{v_i} = (\mathcal{N}^{v_i}, E^{v_i})$ is constructed using these salient objects as nodes $\mathcal{N}^{v_i}$ and encoding the semantic relationship as edge $e_{i,j} \in E^{v_i}$. We have utilized volumetric saliency [6] to detect the salient object and corresponding score $\breve{S}_{O_i}$ as saliency score. The edge (semantic relation) between the salient objects O_i and O_j is computed based upon geometry i.e. spatial distance $d_s(i,j)$ and mean dimensions l_i and l_j of objects O_i and O_j respectively as in 4.

$$E^{v_i}(i,j) = \begin{cases} 1; & \text{if } \min(l_i, l_j) > d_s(i,j) \\ 0; & \text{Otherwise} \end{cases} \tag{4}$$

These undirected graphs of each view are further merged to construct a multi-view summary graph.

B. Multi-view graph construction: We construct a multi-view graph $G = (\mathcal{N}, E, W)$ for a scene by merging all single-view graphs $G^{v_i} = (\mathcal{N}^{v_i}, E^{v_i})$; $i = 1, ..., V$. We fuse all the nodes in $\mathcal{N}$ based on common nodes between views and *Epipolar geometry*. The common nodes are determined based upon the similarity between features which are a concatenation of the object's $\mathcal{O}_k$ multiple features, including location $\mathcal{P}_k^{v_i}$ (estimated via point cloud), saliency score $\breve{S}_{O_i}$, and node labels $L_k^{v_i}$. Thus, the set of fused nodes are given as $\mathcal{N}^m \in \{\mathcal{O}_k | \mathcal{O}_k \in \mathcal{N}^{v_i}, \mathcal{O}_k \in \mathcal{N}^{v_j}$ and $L_k^{v_i} = L_k^{v_j}$ and $\mathbf{C}(\mathcal{P}_k^{v_i}, \tilde{\mathcal{P}}_k^{v_j}) < \tau\}$. Here, $\mathbf{C}$ is the Chamfer distance [20], and $\tau = 1.89$ is the threshold (refer supplementary[1]). Distance coordinate transformation parameters, including rotation R_E and translation t_E are involved in the computation of distance between objects of two views v_i and v_j. We have utilized the method in Efe et al. [21] for finding the set of correspondence points and subsequently estimating the *Epipolar geometry* i.e. translation vector $t_E = [t_x, t_y, t_z]'$ and rotation matrix R_E. The relationship between the locations can be transformed as $\tilde{\mathcal{P}}_k^{v_j} = \mathcal{P}_k^{v_j} R_E + t_E$. The two graphs then merged using these common nodes and keeping edges intact as $E = [E^{v_1} \cup E^{v_2} ... \cup E^{v_V}]$. We further improve the encoding of the semantic information by assigning a weightage $w_{i,j} = W(i,j)$ to the connection between two nodes O_i and O_j as geometric mean of saliency scores $\breve{S}_{O_i}$ and $\breve{S}_{O_j}$ as in 5. Further analysis of the effect of different weights is given in Sec. 5.

$$w_{ij} = \sqrt{\breve{S}_{O_i} \breve{S}_{O_j}} \tag{5}$$

[1] https://sigport.org/documents/indoor-scene-localization-method-using-graphical-summary-multi-view-rgb-d-images

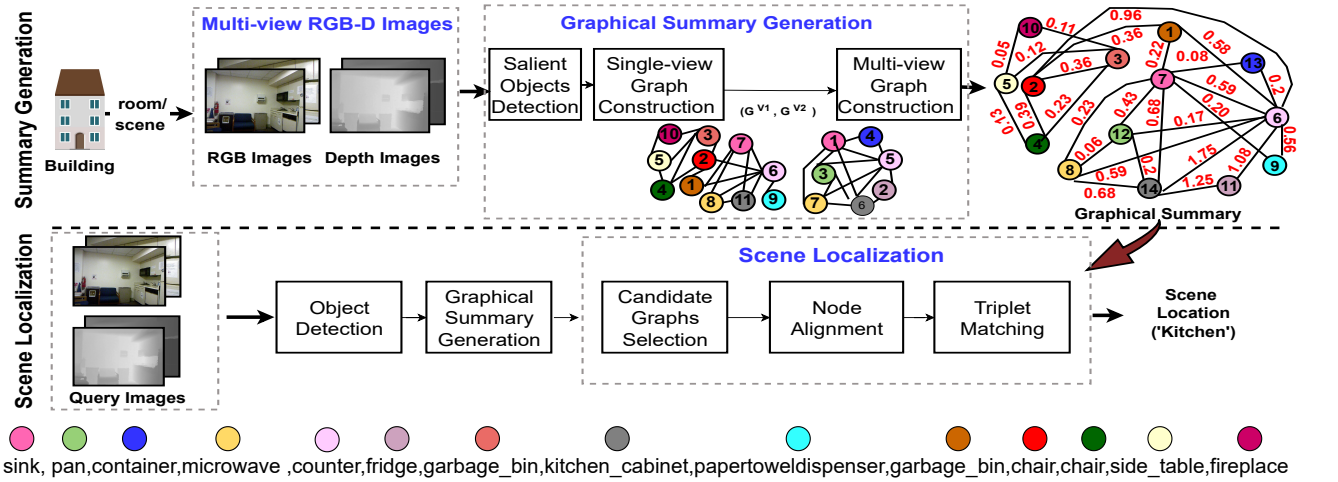

Fig. 1. Framework of the proposed graphical summary generation and scene localization method for multi-view RGB-D images.

4.2. Indoor Scene Localization

This subsection provides details about the proposed method for indoor scene localization via a summary graph-matching approach. An RGB-D image is provided as a query input image. For graph creation, objects in the scene are taken as nodes which are detected using *YOLOv8* [22] trained on *SUNRGB-D* [23]. For this, a pretrained *YOLOv8* [22] is fine-tuned on objects categories of *SUNRGB-D* by splitting data into training, validation, and test sets consisting of 7235, 1550, and 1550 images, respectively. For each detected object $\mathcal{O}_i$, its 3D points (point cloud) data [15] is estimated. Further, we computed the object's 3D bounding box [24]. This information is utilized to create the graphical summary (i.e., query graph G^q) of the input scene as discussed in Sec. 4.1. Given a query graph G^q, the summary graph matching approach aims to identify the best match to G^q in the database $G_i^{\mathcal{D}}$ for predicting the scene $\mathcal{S}^g$. We divide the indoor scene localization approach into three steps: *Candidate Matching*, *Node Alignment*, and *Scene Graph Matching*.

I. Candidate Matching: Given G^q, we first select the possible matched graphs $G_{\hat{i}}^{\mathcal{D}}$ from the constructed graph database $G_i^{\mathcal{D}}$ as in 1. *Jaccard similarity* index $\mathcal{J}_i^q$ between the node's label set L^q of G^q and $L_i^{\mathcal{D}}$ of $G_i^{\mathcal{D}}$ has been utilized. Instead of a strict matching, all the graphs $G_i^{\mathcal{D}}$ with the Jaccard index $\mathcal{J}_i^q$ greater than a threshold is utilized to increase the robustness of the proposed method as in 6.

$$\text{Select all } G_i^{\mathcal{D}} \qquad \mathcal{J}_i^q \geq 0.6 \times \max(\mathcal{J}_i^q) \tag{6}$$

The selected graphs $G_i^{\mathcal{D}}$ are the candidate graphs $G_{\hat{i}}^{\mathcal{D}}$ that are further processed for node alignment to get a correspondence of nodes in $G_i^{\mathcal{D}}$ to the nodes in the query graph G^q.

II. Node Alignment: Node alignment aims to find node matches for every node $\nu_i \in \mathcal{N}^q$ in G^q to a node $\tilde{\nu}_j \in \mathcal{N}^c$ in $G_{\hat{i}}^{\mathcal{D}}$, by finding node-wise similarity metric $\mathbf{d}^{\hat{i}}$. We have utilized the feature vectors f_k and $\tilde{f}_j$ of nodes ν_k and $\tilde{\nu}_j$ respectively for computation of node-wise similarity $\mathbf{d}^{\hat{i}}(f_k, \tilde{f}_j)$. The node feature descriptor f consists of its label, saliency score $\breve{S}_i$, and an embedding vector extracted using the node2vec algorithm [25]. We have considered the normalized projection of feature vector $\tilde{f}_j$ on f_k as the similarity metric. Using this, the alignment matrix $A_{\mathcal{G}^c}$ is computed as given in 2. Then, the candidate subgraph $\mathcal{G}_{\hat{i}}^c$ is extracted from $\mathcal{G}_{\hat{i}}^{\mathcal{D}}$ by retaining only matched nodes. Figure 2(a) shows an example of node alignment between the query and candidate graphs.

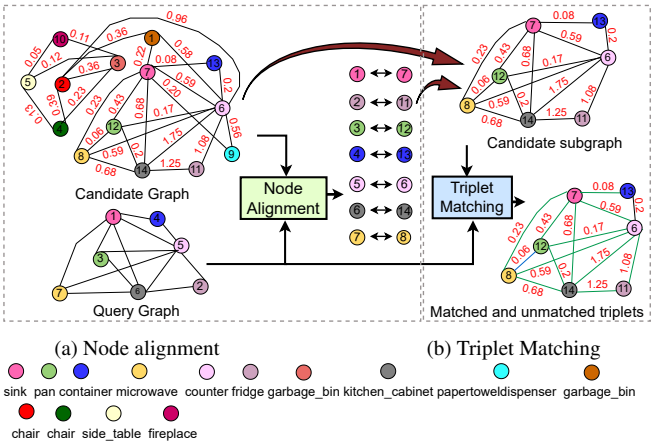

Fig. 2. An example of summary graph matching; matched triplets(green) and unmatched triplets(blue).

III. Scene Graph Matching: Finally, we find out the best matching graph among all the matched candidate graphs $\mathcal{G}_{\hat{i}}^c$ as in 3. Since G^q is a partial subgraph of $\mathcal{G}_{\hat{i}}^c$, exact matching between graphs structure is not possible. Therefore, we employ a triplet matching [26] approach to match the candidate graphs $\mathcal{G}_{\hat{i}}^c$. In this paper, we incorporate edge weights in triplet matching to capture the semantic similarity between all the connections and their importance in a scene graph. The weights W_m and W_u respectively for each of the matched triplet and unmatched triplet are accumulated. During triplet matching, we consider the weights of connection as in 5 irre-

spective of the existence of the edge. The candidate graph $\mathcal{G}_{\hat{i}}^{c}$ is taken as matched only if matched weights W_m are greater than the unmatched weights W_u i.e. $W_m > W_u$. This ensures the removal of the spurious scene matching. Then, the location $\mathcal{S}^g$ is estimated from the candidate graphs $\mathcal{G}_{\hat{i}}^{c}$ with maximum matched weights.

5. RESULTS AND DISCUSSION

This section presents the qualitative, quantitative, and robustness analysis of the proposed graphical summary-based scene localization and compares the performance with state-of-the-art methods.

Dataset Preparation: We have utilized scenes from the SUNRGB-D [23] dataset to create this multi-view graph database of 10 buildings. A total of 224 RGB-D images are included for 10 buildings. The ground truth object labels and corresponding 3D bounding boxes are also given in the SUNRGB-D [23] dataset. These sets of multi-view RGB-D images along with the object information are further utilized to generate single graphical summaries for each scene.

5.1. Quantitative Comparison:

We compare the performance of the proposed method for scene localization in the created dataset for 60 query RGB-D images using *Accuracy*, *F1-Score*, *Precision*, and *Recall* metrics. Table 1 presents mean value of the performance metrics of the proposed method along with the existing methods including, [4], [13], [15], [17], [5], and [18]. We notice from the table that the methods in [4] and [13] which utilize visual features extracted using CNN do not achieve good results in terms of metrics. In a comparison of [4] and [13], the method in [17] improves the value of the metrics by mapping the CNN-based extracted features into high-level representations with a fully randomized structure of recursive neural networks (RNNs). The inclusion of geometric and texture features [15] improves the performance compared to visual features [4] and [13] but fails in capturing object semantic features. Instead of visual features, the method in [5] employed semantic features and object relations for scene localization. This leads to an improvement in metrics value in comparison of [4], [13], [17], and [15]. Recently proposed deep residual learning-based method [18] achieves a higher value of comparison metrics in comparison to others but lower than the proposed method due to its dependency on single-view RGB-D image. However, the proposed method yields maximum performance metrics among all the methods as semantic scene information and relations in intra-view and inter-view of the scene are also utilized unlike in the case of existing methods. Additionally, instead of using an entire scene or all objects within a scene by assigning them equal importance, the proposed method utilizes a saliency-based graphical summary for localization.

In addition, we also demonstrate the significance of each component in the proposed multi-view graphical summary-based localization method by reporting the variation in results due to the presence of these components. We compare our method that includes both nodes (i.e., salient objects) and weighted triplet (semantic information), with the methods that only use either objects and TF-IDF weighting [11], visual feature [27], object and scene features [28], node/object label-based, or unweighted triplet matching. The results of this comparison are shown in Tab. 2 as the mean value of performance metrics over 224 test inputs.

Table 1. Performance comparison of scene localization.

Method	GLN-STA -ATT[4]	TRecgNet [13]	2D-3D Fus-ionNet[15]	ResNet101-RNN[17]	OTS [5]	Omnivore (Swin-L)[18]	Proposed
Input Data	Muli-view RGB	RGB-D	Point cloud and RGB	RGB-D	RGB	RGB-D	Muli-view RGB-D
Accuracy	0.54	0.76	0.77	0.8	0.88	0.93	**0.95**
Precision	0.82	0.87	0.87	0.88	0.90	0.95	**0.97**
Recall	0.62	0.87	0.89	0.89	0.88	**0.97**	**0.97**

Table 2. Performance comparison of scene localization using different scene features (*OL=salient object's label based, TW= unweighted graphical summary-based).

Method	[11]	[27]	OL	[28]	TW	Proposed
Accuracy	0.71	0.75	0.82	0.83	0.84	**0.90**
F1-Score	0.83	0.85	0.89	0.90	0.91	**0.94**
Precision	0.86	0.87	0.88	0.88	0.88	**0.95**
Recall	0.80	0.85	0.92	0.94	**0.95**	0.93

We observe from the table that the method that performs localization based on objects and TF-IDF weighting [11] achieves the lowest performance metrics value in comparison to others as it does not consider the additional object's information e.g. size, geometrical features, etc. The method in [27] utilizes the visual feature extracted for an entire input image and hence does not capture adequate scene information for localization. We also performed the scene localization based on labels of summary i.e. salient objects. Volumetric-based saliency map is utilized to extract the salient objects to generate the summary for localization. Then, the scene identification is performed based on Jaccard similarity between the labels of salient objects of the query image and images in the dataset. We notice from the table that it improves the performance in comparison to [11] and [27] but it lacks the additional scene information. Whereas, the method in [28] employed a network that utilizes both object and scene features which results in the increment in the performance metrics. However, the method still lacks the crucial semantic relationships among the objects in the scene. This semantic information can improve the performance of the scene localization task. Inspired by the work in [28], we incorporated the salient features of the objects and the semantic relationships between them to encode the scene's information that leads to an unweighted graphical summary of a scene (TW). Hence, the metrics values obtained by this unweighted graphical summary-based localization are better in compari-

son to others. However, this approach also does not capture semantic scene information accurately. Thus, in the proposed method, instead of an unweighted graph, we encode the semantic information as a weighted graph which leads to a further significant increment in performance metrics. Thus, the results demonstrate the benefit of considering both salient objects as nodes (visual) and weighted triplet (semantic) for localization.

Table 3. Effects of edge weights on performance using maximum relative score $w_{ij}^{max} = \max(\breve{S}_i\breve{S}_j)$, minimum relative score $w_{ij}^{min} = \min(\breve{S}_i\breve{S}_j)$, average score=$w_{ij}^{avg} = (\breve{S}_i + \breve{S}_j)/2$, inverse score $w_{ij}^{inv} = \breve{S}_i^{-1} + \breve{S}_j^{-1}$ and proposed $w_{ij}^{Pro} = \sqrt{\breve{S}_i\breve{S}_j}$.

Edge weights	w_{ij}^{max}	w_{ij}^{min}	w_{ij}^{avg}	w_{ij}^{inv}	w_{ij}^{Pro}
Accuracy	0.83	0.85	0.88	0.75	**0.90**
F1-Score	0.89	0.91	0.93	0.84	**0.94**
Precision	0.91	0.93	**0.95**	0.87	**0.95**
Recall	0.92	0.91	0.92	0.84	**0.93**

Furthermore, to demonstrate the effectiveness of edge weights w_{ij}^{Pro} considered in the proposed method, we performed localization using different edge weights including, w_{ij}^{max}, w_{ij}^{min}, w_{ij}^{avg}, w_{ij}^{inv}, and w_{ij}^{Pro}. The results of this comparison are shown in Tab. 3 as the mean value of performance metrics over 224 test inputs. We observe from the table that the scene identification performed using w_{ij}^{Pro} outperforms others. Although the metric *Precision* in case of w_{ij}^{avg} achieve the same value, its performance is slightly lesser than w_{ij}^{Pro} in terms of the remaining three metrics (i.e., *Accuracy*, *F1-Score*, and *Recall*).

5.2. Qualitative Comparison:

We also present the qualitative comparison between the proposed multi-view graphical summary-based scene localization method and methods proposed in [4], [13], [17], [15], [5], and [18]. Figure 3 shows the comparison results for a few test images obtained using SOTA methods and the proposed method. The first two rows depict the created multi-view dataset for two sample buildings containing 6 and 4 scenes. The table below shows some sample results of scene localization for test inputs for this dataset. The ground truth label *BiSj* denotes the input taken from the j_{th} scene of the i_{th} building. The figure shows that the visual features-based methods [4], [13], and [17] fail in the case of the scene contains objects that have visual similarity like color. Also, these methods do not perform well under unconditional environment conditions like illumination, etc. The predicted labels using geometric and texture features-based method [15], semantic features and object relations-based method [5] mislead in the case of the scene having similar spatial arrangements due its dependency on the single view that results in the absence of inter-relation among multi-views of the complete scene. Multi-modality features-based method [18] performs better than other existing methods but it also predicts the location of a scene that is not present in the building. In contrast, the proposed method outperforms others as it considers the salient object as nodes (contextual information) and semantic relationship in intra-view and inter-view. However, the method in [4] also utilizes multi-view information but due to assigning equal importance to all objects/components in a scene it does not perform well.

Table 4. Comparison of accuracy of proposed method under noise, blur+noise, color, geometric artifacts. (OL =salient object's label-based)

Artifact	Parameters	Image-based	OL	Proposed
Noise	$SNR = 21$	0.75	0.82	**0.90**
	$SNR = 16$	0.44	0.82	**0.87**
	$SNR = 10$	0.22	0.73	**0.81**
	$SNR = 4.5$	0.13	0.66	**0.69**
Blur+ Noise	$\sigma_b = 2$ & $SNR = 21$	0.11	0.57	**0.63**
	$\sigma_b = 2$ & $SNR = 16$	0.04	0.50	**0.54**
	$\sigma_b = 3$ & $SNR = 16$	0.02	**0.34**	0.31
Color Contrast	$\gamma = 0.5$	**0.68**	0.57	0.63
	$\gamma = 1.5$	0.66	0.69	**0.71**
	$\gamma = 2.0$	0.66	0.66	**0.71**
Rotation	$R = 10°$	0.68	0.82	**0.90**
	$R = 30°$	0.66	0.80	**0.88**

5.3. Robustness Analysis:

We have also examined the performance of the proposed graphical summary-based scene localization method under different types of artifacts. We have examined the performance under four types of artifacts viz. noise, blur+noise, color, and geometric. We utilized the trained *YOLOv8* [22] (see Sec. 4.2) for object detection and then extracted salient objects from all detected objects using [6]. Then, the scene identification is performed based on Jaccard similarity between the labels of salient objects of the query image and images in the dataset. The variations in *Accuracy* observed for visual features-based, salient objects-based, and summary-based scene identification due to the introduction of various artifacts is reported in Tab. 4. We notice from the table that the performance of all three (i.e., visual features-based, salient objects-based, and summary-based) approaches reduced with increasing the amount of distortion. Because of the limitation of visual features in an unconditional environment, the visual features achieve the lowest metric values in comparison to the other two approaches for all cases except for $\gamma = 0.5$. However, the proposed method not only enhances the overall scene localization performance but also increases the robustness of the method against artifacts. Thus, the results demonstrate the effectiveness of the proposed method under various unconditional environment conditions.

6. CONCLUSION

This paper presents a novel method for indoor-scene localization using the graphical summary from multi-view RGB-D images. Additionally, the paper also presents a novel method for multi-view graphical summary generation. The proposed

Query Images	Ground truth scene labels	Predicted scene labels						
		GLN-STA-ATT[4]	TRecgNet [13]	ResNet101 -RNN[17]	2D-3D FusionNet[15]	OTS [5]	Omnivore (Swin-L)[18]	Proposed
	B1S1	B1S4	B1S4	B1S1	B1S6	B1S1	B1S1	B1S1
	B1S1	B1S4	B1S2	B1S4	B1S4	B1S4	B1S1	B1S1
	B1S3	B1S4	B1S3	B1S5	B1S6	B1S5	B1S3	B1S3
	Not present	B1S5	B1S3	B1S5	B1S3	B1S5	B1S3	Not present
	B1S5	B1S5	B1S3	B1S3	B1S2	B1S6	B1S5	B1S5
	Not present	B1S1	B1S2	B1S4	B1S4	B1S1	B1S1	B1S1
	B1S6	B1S1	B1S4	B1S6	B1S6	B1S1	B1S4	B1S6
	B2S1	B2S1	B2S2	B2S3	B2S2	B2S1	B2S1	B2S1

Fig. 3. Qualitative comparison of scene localization performance. Top two rows: example multi-view dataset for two buildings. Third to bottom rows: query images with ground truth scene label and predicted scene labels using methods in [4], [13], [17], [15], [5], and [18] and proposed method, (B=building, S=scene).

method utilizes visual and geometrical features along with semantic features for summary graph generation. Semantic information is encoded in terms of weighted connection based on the saliency value. Furthermore, this work utilizes a graph-matching approach that utilizes both node-wise and edge-wise similarity for scene localization. The results demonstrate the efficacy of the proposed multi-view graphical summary-based method for indoor scene localization. However, the proposed method shows limited performance in the case of single-view images containing a few objects that are also present in other scenes. The efficacy of the proposed method in dynamic environments has yet to be verified. In the future, we would like to extend the proposed method for encoding the semantic information of the scene using multimodal scene information and to facilitate the localization of outdoor scenes.

7. REFERENCES

[1] Faheem Zafari, Athanasios Gkelias, and Kin K Leung, "A survey of indoor localization systems and technologies," *IEEE Communications Surveys & Tutorials*, vol. 21, no. 3, pp. 2568–2599, 2019. 1

[2] Zhaozheng Hu, Gang Huang, Yuezhi Hu, and Zhe Yang, "Wi-vi fingerprint: Wifi and vision integrated fingerprint for smartphone-based indoor self-localization," in *ICIP*. IEEE, 2017, pp. 4402–4406. 1

[3] Duncan P Robertson and Roberto Cipolla, "An image-based system for urban navigation.," in *BMVC*, 2004, vol. 19, p. 165. 1

[4] Meng-Jiun Chiou, Zhenguang Liu, Yifang Yin, An-An Liu, and Roger Zimmermann, "Zero-shot multi-view

indoor localization via graph location networks," in *ACMMM*, 2020, pp. 3431–3440. 1, 4, 5, 6

[5] Bo Miao, Liguang Zhou, Ajmal Saeed Mian, Tin Lun Lam, and Yangsheng Xu, "Object-to-scene: Learning to transfer object knowledge to indoor scene recognition," in *IROS*. IEEE, 2021, pp. 2069–2075. 1, 4, 5, 6

[6] Preeti Meena, Himanshu Kumar, and Sandeep Yadav, "A volumetric saliency guided image summarization for rgb-d indoor scene classification," *IEEE Transactions on Circuits and Systems for Video Technology*, pp. 1–1, 2024. 1, 2, 5

[7] Guoyu Lu, Yan Yan, Nicu Sebe, and Chandra Kambhamettu, "Indoor localization via multi-view images and videos," *Computer Vision and Image Understanding*, vol. 161, pp. 145–160, 2017. 1

[8] Sheng Han, Wei Gao, Yiming Wan, and Yihong Wu, "Scene-unified image translation for visual localization," in *ICIP*. IEEE, 2020, pp. 2266–2270. 1

[9] Songxiang Yang, Lin Ma, Shuang Jia, and Danyang Qin, "An improved vision-based indoor positioning method," *IEEE Access*, vol. 8, pp. 26941–26949, 2020. 1

[10] Zhitong Xiong, Yuan Yuan, and Qi Wang, "Ask: Adaptively selecting key local features for rgb-d scene recognition," *TIP*, vol. 30, pp. 2722–2733, 2021. 1

[11] Edvard Heikel and Leonardo Espinosa-Leal, "Indoor scene recognition via object detection and tf-idf," *Journal of Imaging*, vol. 8, no. 8, pp. 209, 2022. 1, 4

[12] Xinhang Song, Shuqiang Jiang, Bohan Wang, Chengpeng Chen, and Gongwei Chen, "Image representations with spatial object-to-object relations for rgb-d scene recognition," *TIP*, vol. 29, pp. 525–537, 2019. 1

[13] Dapeng Du, Limin Wang, Zhaoyang Li, and Gangshan Wu, "Cross-modal pyramid translation for rgb-d scene recognition," *IJCV*, vol. 129, no. 8, pp. 2309–2327, 2021. 1, 4, 5, 6

[14] Liguang Zhou, Jun Cen, Xingchao Wang, Zhenglong Sun, Tin Lun Lam, and Yangsheng Xu, "Borm: Bayesian object relation model for indoor scene recognition," in *IROS*. IEEE, 2021, pp. 39–46. 1

[15] Albert Mosella-Montoro and Javier Ruiz-Hidalgo, "2d–3d geometric fusion network using multi-neighbourhood graph convolution for rgb-d indoor scene classification," *Information Fusion*, vol. 76, pp. 46–54, 2021. 1, 3, 4, 5, 6

[16] Boney A Labinghisa and Dong Myung Lee, "Indoor localization system using deep learning based scene recognition," *Multimedia Tools and Applications*, vol. 81, no. 20, pp. 28405–28429, 2022. 1

[17] Ali Caglayan, Nevrez Imamoglu, Ahmet Burak Can, and Ryosuke Nakamura, "When cnns meet random rnns: Towards multi-level analysis for rgb-d object and scene recognition," *Computer Vision and Image Understanding*, vol. 217, pp. 103373, 2022. 1, 4, 5, 6

[18] Rohit Girdhar, Mannat Singh, Nikhila Ravi, Laurens van der Maaten, Armand Joulin, and Ishan Misra, "Omnivore: A single model for many visual modalities," in *CVPR*, 2022, pp. 16102–16112. 2, 4, 5, 6

[19] Ricardo Pereira, Tiago Barros, Luís Garrote, Ana Lopes, and Urbano J Nunes, "A deep learning-based global and segmentation-based semantic feature fusion approach for indoor scene classification," *Pattern Recognition Letters*, 2024. 2

[20] Trung Nguyen, Quang-Hieu Pham, Tam Le, Tung Pham, Nhat Ho, and Binh-Son Hua, "Point-set distances for learning representations of 3d point clouds," in *ICCV*, 2021, pp. 10478–10487. 2

[21] Ufuk Efe, Kutalmis Gokalp Ince, and Aydin Alatan, "Dfm: A performance baseline for deep feature matching," in *CVPR*, 2021, pp. 4284–4293. 2

[22] Armstrong Aboah, Bin Wang, Ulas Bagci, and Yaw Adu-Gyamfi, "Real-time multi-class helmet violation detection using few-shot data sampling technique and yolov8," in *CVPR*, 2023, pp. 5349–5357. 3, 5

[23] Shuran Song, Samuel P Lichtenberg, and Jianxiong Xiao, "Sun rgb-d: A rgb-d scene understanding benchmark suite," in *CVPR*, 2015, pp. 567–576. 3, 4

[24] Chia-Tche Chang, Bastien Gorissen, and Samuel Melchior, "Fast oriented bounding box optimization on the rotation group so (3,r)," *ACM Transactions on Graphics (TOG)*, vol. 30, no. 5, pp. 1–16, 2011. 3

[25] Aditya Grover and Jure Leskovec, "node2vec: Scalable feature learning for networks," in *Proceedings of the 22nd ACM SIGKDD international conference on Knowledge discovery and data mining*, 2016, pp. 855–864. 3

[26] Zhenguang Liu, Li Cheng, Anan Liu, Luming Zhang, Xiangnan He, and Roger Zimmermann, "Multiview and multimodal pervasive indoor localization," in *ACMMM*, 2017, pp. 109–117. 3

[27] S Jian, H Kaiming, R Shaoqing, and Z Xiangyu, "Deep residual learning for image recognition," in *CVPR*, 2016, pp. 770–778. 4

[28] Hongje Seong, Junhyuk Hyun, and Euntai Kim, "Fosnet: An end-to-end trainable deep neural network for scene recognition," *IEEE Access*, vol. 8, pp. 82066–82077, 2020. 4