Vector Symbolic Scene Representation for Semantic Place Recognition

Daniil Kirilenko
Moscow Institute of Physics and Technology
Dolgoprudny, Russia
kirilenko.de@phystech.edu

Alexey K. Kovalev
AIRI
Moscow, Russia
kovalev@airi.net
HSE University
Moscow, Russia

Yaroslav Solomentsev
Moscow Institute of Physics and Technology
Dolgoprudny, Russia
solomentsev.yak@phystech.edu

Alexander Melekhin
Moscow Institute of Physics and Technology
Dolgoprudny, Russia
melekhin.aa@mipt.ru

Dmitry A. Yudin
Moscow Institute of Physics and Technology
Dolgoprudny, Russia
yudin.da@mipt.ru
AIRI
Moscow, Russia

Aleksandr I. Panov
AIRI
Moscow, Russia
panov@airi.net
FRC CSC RAS
Moscow, Russia

***Abstract*—Most state-of-the-art methods do not explicitly use scene semantics for place recognition by the images. We address this problem and propose a new two-stage approach referred to as TSVLoc. It solves the place recognition task as the image retrieval problem and enriches any well-known method. In the first model-agnostic stage, any modern neural network model that does not directly use semantics, e.g., HF-Net, NetVLAD, or Patch-NetVLAD, can be used. In the second stage, we apply the Vector Symbolic Architectures (VSA) framework to construct semantic scene representation. Our method uses semantic segmentation of an image to extract objects and their relations and applies VSA operations to form semantic scene representation. For this, an optional usage of the depth map was considered, which showed promising results. The effectiveness of our approach is demonstrated through extensive experiments on the open large-scale datasets: the indoor HPointLoc dataset built in the Habitat simulation environment and the outdoor Oxford RobotCar dataset. The proposed solution significantly improves the quality of the place recognition.**

***Index Terms*—place recognition, image retrieval, vector symbolic architectures, semantic scene representation**

I. Introduction

The solution to the place recognition problem is an essential part of approaches to the global localization of intelligent agents, particularly, robots. The use of onboard camera images for this purpose makes such solutions simpler and cheaper. Furthermore, the advent of affordable RGB-D cameras and fast and high-quality algorithms for semantic segmentation makes it possible to use information about the objects' presence and spatial relationships with various semantic categories.

Such information is typically encoded implicitly in modern image retrieval methods that form a global image embedding at the output. An important modern research focus area is the improvement of such basic methods based on a multi-stage approach to refine the results of matching the query image and the most similar images from the database [1]. However, the straightforward application of semantic information to form informative and interpretable image embeddings is still poorly understood.

This work was supported by the Russian Science Foundation, project No. 20-71-10116.

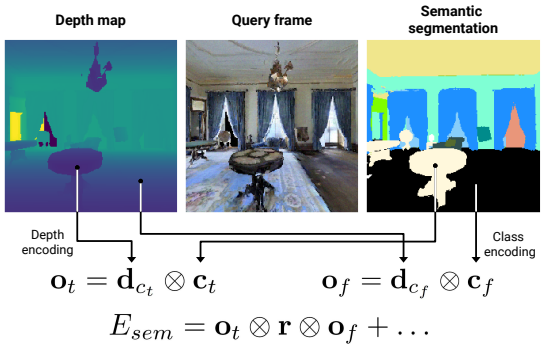

Fig. 1. To form a semantic scene encoding E_{sem}, we use depth and semantic maps of a query frame. For every class C, we generate a high-dimensional vector $\mathbf{c}$, then bind it with a depth vector $\mathbf{d_c}$ of the center of mass of the class instance, and get an object vector $\mathbf{o}$. After that, every pair of vectors that have a common border is bound together through an auxiliary vector $\mathbf{r}$ (relation "near") and summed up. In the example above, $\mathbf{c_t}$ and $\mathbf{c_f}$ are high-dimensional vectors for classes *table* and *floor*; $\mathbf{d_{c_t}}$, $\mathbf{d_{c_f}}$ and $\mathbf{o_t}$, $\mathbf{o_f}$ are depth and object vectors correspondingly

This paper proposes a novel two-stage approach termed TSVLoc, which uses the Vector Symbolic Architectures (VSA) framework [2] to construct semantic scene representation based on input semantic and depth maps. It is designed to enhance the quality of any basic RGB image-based approaches to place recognition.

Vector Symbolic Architectures (VSA) were first proposed

in cognitive psychology and cognitive neuroscience [3] as computational models of the cognitive process. Under this framework, symbols are represented as vectors with a high yet fixed dimension. The symbolic operations, such as assigning a value to an attribute, are performed by vector operations. Thus, it bridges the gap between symbolic and subsymbolic (connectionist) representation. The vectors could be of different nature: binary [4], real [5], or complex [6]. Vector Symbolic Architectures enable us to encode complex structures such as sequences [7], [8], graphs [4], [9], binary trees [10], and even finite-state automata [11] into high-dimensional vectors. In our work, we use this property to encode query frame semantic information, i.e., objects and their relative depth to the camera.

Our contributions are summarized as follows.

- We propose a novel two-stage approach for place recognition referred to as TSVLoc. The first stage of the proposed method is model agnostic, so it has a potential to be used to boost the performance of any traditional method that outputs global embeddings.
- We use the semantic embedding of a scene as an additional frame embedding to refine place recognition results. We construct this embedding from semantic and depth maps of the scene using Vector Symbolic Architectures.
- Extensive experiments have shown that TSVLoc demonstrates a significant improvement for basic methods based on the popular neural network models HF-Net and NetVLAD.

II. Related Work

A. Place Recognition

Image retrieval is a key task for the place recognition problem. It involves finding and ranking the most similar images from the database to a query image. Its solution requires generating informative and compact local and global descriptors of the images.

The common "classical" approaches obtain global features (embeddings) by aggregating the local descriptors using the bag of words (BoW) scheme, e.g., DBoW2, DBoW3 [12], or FBoW [13], or vectors of locally aggregated descriptors (VLAD) [14].

Over the last few years, new approaches based on deep neural networks have been released: NetVLAD model [15], its distilled version HF-Net [16], Ap-Gem approach with differentiable rank loss function [17], etc. These have surpassed the classical ones by feature learning for the specific problems.

The similar image candidates can be also re-ranked by analyzing statistics of geometrically correct matching of local descriptors for image patches as in Patch-NetVLAD [1].

Such approaches can be classified as two-stage, providing a state-of-the-art level of quality when solving the place recognition problem.

To improve the matching of embeddings, some approaches utilize semantic information for global feature vector generation. The key idea of the Vector semantic representations (VSR) approach [18] is to describe the relations between the semantic objects. For example, there is a sidewalk to the right of the street and grass terrain to the left, which, in turn, is followed by another sidewalk and a fence. The semantic ranking method [19] ranks the 2D–3D matches found by the feature-based localization pipeline depending on how well they agree with the scene semantics. Its authors proposed an original approach to learning 3D semantic descriptors. However, in spite of notable achievements of these approaches, there still is a problem with the extraction of invariant and informative semantic descriptors of images in dynamically changing scenes.

B. Semantic Scene Representation

The idea of using semantic information of a scene for image retrieval is being actively developed. In [20], the detailed semantic of the scene is represented by a scene graph. It encodes objects, their attributes, and relationships between objects. The scene graph describes an image in general. To represent a specific image, a scene graph is grounded to it by associating each object with a corresponding region of the image (bounding box).

The problem of generating scene graphs from images is well-studied [21]–[24]. Most of the models rely on message passing [25] between objects or corresponding relations [22], [26]–[29]. Such representation is also applicable for 3D scenes [30], [31]. Some models use an external knowledge base to refine features of extracted objects by commonsense relationships [32].

All of these approaches require learning on a specific dataset (e.g., Visual Genome [33], Open Images [34]) that takes time and resources. In our method, we utilize an image segmentation map available for an agent navigation system and construct a semantic scene representation without any learning. Using this representation for place recognition in a second stage on top of traditional methods leads to performance improvements.

III. Methodology

A. Task Statement

In this paper, we solve the problem of place recognition and estimate the position P_q with three degrees of freedom (x, y, z) of an intelligent agent from the image I_q of its RGB-D camera (query) in the vicinity of the poses of the cameras P^i_{db} from the database, $i \in [1, N]$. The N camera poses from the database form a regular grid of groups of six cameras, the images I^i_{db} of which cover 360° view in the environment.

As the desired pose for the query image, we use the pose of the most similar image from the database with the index i_{top_1}:

$$P_q = P_{db}^{i_{top_1}}.$$

To select the most similar image, we explore both popular image retrieval approaches based on neural networks and the capabilities of Vector Symbolic Architectures for efficiently encoding R_q semantic maps and D_q depth maps as high-dimensional vectors.

This formulation of the visual place recognition problem allows us to investigate the possibilities and quality of an intelligent agent re-localization on previously seen scenes using important additional semantic information.

B. Place Recognition

In the first stage, we can vary several neural network-based methods of place recognition for generating global feature vectors (embeddings) for the query image E_q and images from the database E^i_{db}. We use the popular NetVLAD approach, which generates a vector of size 32,768 elements, as well as its distilled version from the HF-Net approach, which generates a vector of length 4,096. Based on these embeddings, the top-n similar image (with index $i^*_{top_n}$) in the sense of the similarity metric S is determined:

$$i^*_{top_n} = \text{topk}\left(S\left(E_q, E^i_{db}\right), n\right),$$

where topk is the operator for selection $k = n$ image indices from database in descending order of the similarity metric S. In this paper, S is the cosine proximity function, the value of which in the range [0, 1] indicates the similarity of images.

C. Semantic Representation

Vector Symbolic Architectures (VSAs) [2], or Hyperdimensional Computing, is a computing framework that operates on high-dimensional vectors. VSAs originate in cognitive psychology and cognitive neuroscience as a connectionist model capable of performing symbolic reasoning. Under this framework, random high-dimensional vectors, or HD vectors, represent symbols, and, consequently, manipulation with them reduces to vector operations. The randomness of HD vectors means that to encode a basic symbol, we sample a vector from a vector space of dimension D (typically D is greater than 1,000) and use it as a seed hypervector. VSAs distribute the encoded information across all components of the HD vector. Thus, only the whole vector could be interpreted. Distributed representation is different from localist representation, where every single vector component has a meaning. The concentration of the measure phenomenon [35], [36] ensures that random vectors from a high-dimensional vector space are almost orthogonal (quasi-orthogonal), and thus representations of different basic symbols are dissimilar. The nature of the vector space might be different that results in binary, real, or complex HD vectors. These seed hypervectors are stored in the item memory, from which they could be extracted and used to form complex structures as composite vectors. To encode complex structures into HD vectors, VSAs offer several vector operations. The exact implementation of vector operations varies for different vector spaces while keeping computational properties. We explain these operations using an example of the Multiply-Add-Permute [5] variation of VSAs that works with real vectors.

The similarity measure is the basis for reasoning in VSAs, as it is used to extract seed hypervectors from the item memory and compare complex HD vectors.

The addition operation or bundling (denoted as $+$) is an element-wise sum. The resultant vector is similar to summand vectors but quasi-orthogonal to others. Bundling represents a set of vectors and, correspondingly, a set of symbols.

The multiplication operation or binding (denoted as $\odot$) is an element-wise multiplication of two HD vectors. It maps these vectors to another HD vector. The resultant hypervector is dissimilar (quasi-orthogonal) to multiplied and other HD vectors from the vector space. Semantically, binding represents an attribute-value pair, an assignment of a value to a corresponding attribute. Binding and bundling are core operations for encoding complex structures into HD vectors.

In this paper, we use a semantic scene representation inspired by [37]–[39]. We consider the scene as a collection of objects, their properties, and relations between them. The semantic map is used to detect objects presented on the scene. For each object, an HD vector is generated. If objects belong to the same class, we use the same vectors to encode them. The only relation between objects we use is a "near" relation. It indicates that one object is close to another, and, on the semantic map, this is expressed in the fact that two corresponding segments have a common border.

The previous works [37]–[39] used simple schemes to encode location and did not handle real-valued coordinates. Therefore, as VSA, we use a variance of the Semantic Pointer Architecture (SPA) [6], Spatial Semantic Pointers [40], as it offers various schemes for encoding structured continuous space. Real random vectors with a unit norm are used as seed hypervectors. The binding operation in SPA is a circular convolution (denoted as $\otimes$):

$$\mathbf{u} \otimes \mathbf{w} := IDFT(DFT(\mathbf{u}) \odot DFT(\mathbf{w})),$$

where $\mathbf{u}, \mathbf{w}$ are two HD vectors; DFT and IDFT denote the Discrete Fourier Transform and Inverse Discrete Fourier Transform accordingly.

SPA also defines the fractional binding operation:

$$\mathbf{u}^p := \Re(IDFT((DFT(\mathbf{u})^p)_{j=0}^{D-1})),$$

where $\Re$ denotes the real part of a number.

These operations enable us to encode numerical values corresponding to the coordinates x, y by generating two unitary HD vectors for the coordinate axes $\mathbf{X}, \mathbf{Y}$ (vector $\mathbf{u}$ is called unitary if $\forall \mathbf{v} : \|\mathbf{v}\| = \|\mathbf{v} \otimes \mathbf{u}\|$) and applying fractional binding:

$$\mathbf{V} = \mathbf{X}^x \otimes \mathbf{Y}^y.$$

In our work, this is used to encode the depth of objects relative to the camera in range $[0, 2]$.

We encode semantic information into an HD vector E_{sem} as follows. The class extraction function CE outputs a list of class instances L_C for a semantic map R_q: $CE(R_q) = L_c$. For every class C from L_C, we generate an HD vector $\mathbf{c}$ (sample from $\mathbb{R}^D$). The depth extraction function DE takes a depth map D_q and L_C. For every class instance from L_C, DE outputs a relative depth from the camera to the instance encoded with a fractional binding as a vector $\mathbf{d}^\ell_c$, where ℓ

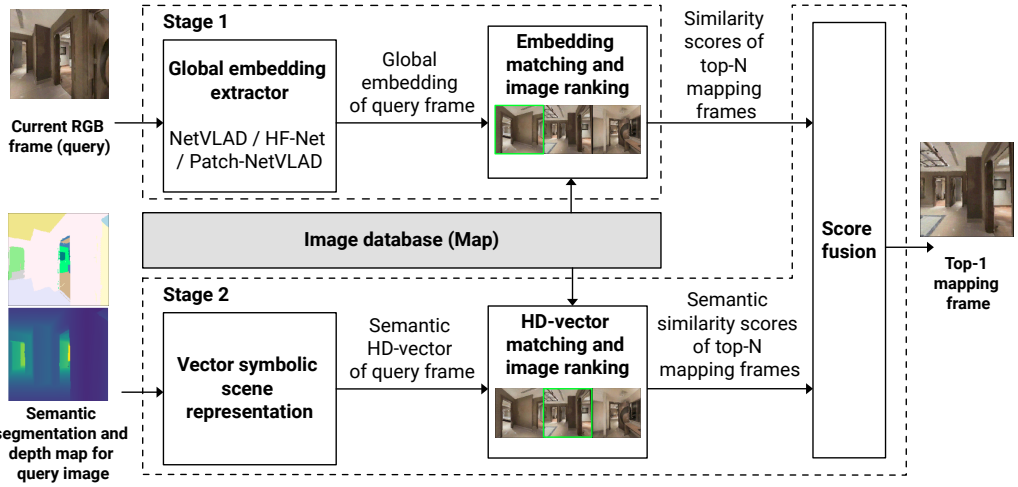

Fig. 2. Overview of the proposed model. The model involves two stages. In the first stage, a global embedding is extracted using one of the traditional methods. Then, an image database is queried, and a ranking of images is produced. In the second stage, the semantic HD vector of a query image is constructed from semantic segmentation. We use Vector Symbolic Architecture to form a semantic vector. Next, we query an image database and get the image ranking for semantic vectors. After that, the current ranking scores, together with the first-stage scores, are passed on to the score fusion module to produce a top-1 ranking image.

is an instance number. All vectors are stored in the list L_d: $DE(D_q, L_c) = L_d$. We bind an instance class vector $\mathbf{c}$ with an instance depth vector $\mathbf{d}_c^\ell$ to get an HD vector $\mathbf{o}$ for the object: $\mathbf{c} \otimes \mathbf{d}_c^\ell$. To construct a final vector E_{sem}, we find all object pairs on a semantic map that have a common border, encode these pairs as $\mathbf{o}_i \otimes \mathbf{r} \otimes \mathbf{o}_j$, where an auxiliary seed HD vector $\mathbf{r}$ represents relation "near", and sum them up. Thus, the semantic vector is:

$$E_{sem} = \sum_{i,j} \left(\mathbf{o}_i \otimes \mathbf{r} \otimes \mathbf{o}_j\right),$$

where objects i, j have a common border (Fig. 1).

The most relevant to our approach is the work in [18]. The authors use a representation called Vector Semantic Representation (VSR), which is obtained by a combination of HD vectors and feature map responses of salient regions from DELF [41]. Meanwhile, in our model, there is no mixing of embeddings of different approaches. Also, for combining VSR and global embeddings, an element-wise multiplication of their pairwise image similarity matrices is computed. We use the two-stage re-ranking approach described in the next section.

D. Proposed Method Structure

In this paper, we propose a two-staged model (Fig. 2) for a place recognition task with a semantic scene representation by Vector Symbolic Architecture. We called this model TSVLoc for Two-Staged VSA Localization.

The first stage of the proposed architecture is model-agnostic. Thus, any traditional method (HF-Net [16], NetVLAD [15], etc.) could be used. Given an RGB-D image, a global embedding E_q is produced. Then, the image database is queried to get top-n similar images $i^*_{top_n}$ and their scores.

In the second stage, the semantic segmentation map and depth information is used to construct a semantic scene representation E_{sem}. Then, semantic similarity scores S_{sem} for every $i^*_{top_n}$ are calculated. After that, we scan through $i^*_{top_n}$: the first candidate image from $i^*_{top_n}$ is considered as a result. We move to the next image. If the similarity S decreases by less than γ_1 relative to the current result and the semantic similarity S_{sem} increases by more than γ_2, this image is considered as a result, where γ_1 and γ_2 are tunable hyperparameters.

E. Quality Evaluation

To evaluate the localization quality of the proposed method, we use the Recall (R) metric with different thresholds. It is calculated as the fraction of query images whose translation errors do not exceed the specified thresholds $\epsilon_t \in \{0.5m, 1m, 5m, 10m\}$. Such thresholds were chosen to assess the accuracy of solving the global indoor localization problem at various spatial scales. We do not take into account the rotation error because we do not optimize the camera pose after image retrieval.

IV. Experiments

A. Datasets

The paper considers the problem of place recognition both in an indoor and outdoor environments. The results are planned to be used in the navigation of intelligent agents.

A flexible tool for carrying out indoor experiments is the photorealistic Habitat simulator [42]. However, it contains only 3D models of various rooms and an interface for extracting data (images, semantic maps, depth maps, poses). To investigate the quality of semantic image retrieval as one of the key stages of place recognition, we chose the open dataset

HPointLoc[1], created on the basis of this simulator with the Matterport3D environment [43].

The HPointLoc dataset contains 86,678 RGB-D images with a resolution of 256×256. Ground truth segmentation maps include 41 semantic categories for 488,717 labeled objects. Its peculiarity is the representation of key camera poses in the form of a regular grid that covers the 3D scene.

To demonstrate more clearly the specific impact of our semantic approach on the place recognition results, we defined a HPointLoc subset with query images difficult for basic image retrieval method. We selected queries for which the localization error, when using the HF-Net method, is greater than 0.5m. This subset contains 9,324 query images (see Fig. 3). We termed this subset HPointLoc-Hard.

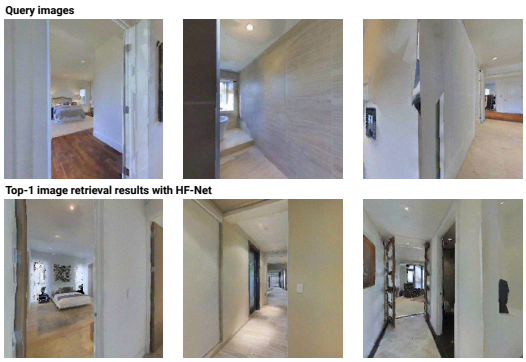

Fig. 3. Examples of query images and HF-Net image retrieval results from a prepared subsample of the HPointLoc dataset termed HPointLoc-Hard. We can see that the Top-1 image selection is wrong for these cases

In addition to the HPointLoc dataset, we validated our approach on subsamples of the outdoor Oxford RobotCar dataset [44]. We took the same subsamples as in [18]: "2014-11-25-09-18-32" (2,244 images), "2014-12-09-13-21-02" (2,210 images), "2014-12-16-18-44-24" (1,916 images), "2015-02-03-08-45-10" (2,408 images), "2015-05-19-14-06-38" (2,065 images), and "2015-08-28-09-50-22" (2,072 images). We took only the central stereo camera data from each subsample. The resolution of the images is 1,280×960. Also, we did not use all the images from the camera. We found the first frame closest to zero ground truth timestamp. Next, we sampled frames with a frequency of 1 Hz and selected the poses closest to them in terms of timestamp.

Semantic segmentation for the Oxford RobotCar dataset was performed using the HRNet + OCR[2] model trained on Mapillary Vistas dataset [45].

B. Place Recognition Performance

To estimate the place recognition quality, we use the Recall localization metric on all query images with different distance thresholds.

Results on the indoor HPointLoc and outdoor the Oxford RobotCar datasets for different image retrieval methods are shown in Table I and Table II correspondingly. We do not use SEBD mode for the Oxford RobotCar dataset since the depth map is not available.

During the experiments, we used real vectors with a unit norm and dimension 1,000 for semantic representation. In a two-stage place recognition process, the same hyperparameters were used for all maps from the dataset: γ_1 = 0.06, γ_2 = 0.4, N = 5. The best results were obtained with complex semantic encoding of the scene, which, among other things, takes into account the depth map, encoding the distance to the center of mass of a particular object.

As a basis for our approach, we used three methods: HF-Net[3], NetVLAD[4], and the state-of-the-art method Patch-NetVLAD[5]. TSVLoc(H), TSVLoc(V), and TSVLoc(P) stand for different models used in the first stage, HF-Net, NetVLAD, and Patch-NetVLAD correspondingly. As you can see from Table I, our approach wins at large distances and trails to it at short distances. In experiments on Oxford RobotCar (Table II), our approach with Patch-NetVLAD outperforms all other approaches.

TABLE I
LOCALIZATION METRICS ON ALL QUERY IMAGES FROM THE HPOINTLOC DATASET. R(0.5) MEANS THE RECALL METRIC WITH 0.5M DISTANCE THRESHOLD.

Method	R(0.5)	R(1)	R(5)	R(10)
HF-Net	0.890	0.892	0.963	0.976
TSVLoc(H+SEB)	0.892	0.893	0.977	0.988
TSVLoc(H+SEBD)	**0.895**	**0.896**	**0.980**	**0.993**
NetVLAD	0.887	0.888	0.962	0.973
TSVLoc(V+SEBD)	**0.892**	**0.893**	**0.976**	**0.987**
Patch-NetVLAD	**0.942**	0.943	0.968	0.978
TSVLoc(P+SEBD)	0.931	**0.946**	**0.978**	**0.982**

TABLE II
AVERAGED LOCALIZATION METRICS ON THE OXFORD ROBOTCAR DATASET.

Method	R(5)	R(10)	R(25)	R(50)	R(100)
HF-Net	0.485	0.639	0.708	0.737	0.761
TSVLoc(H+SEB)	**0.494**	**0.647**	**0.715**	**0.741**	**0.765**
NetVLAD	0.568	0.725	0.779	0.802	0.822
TSVLoc(V+SEB)	**0.573**	**0.731**	**0.783**	**0.805**	**0.824**
Patch-NetVLAD	0.702	0.842	0.877	0.888	0.898
TSVLoc(P+SEB)	**0.714**	**0.853**	**0.886**	**0.896**	**0.905**

We calculated the localization quality metrics for the HPointLoc-Hard dataset subsample (see Table III). Our two-stage approach does not always improve the response of the original model; sometimes it turns out that due to the two-stage approach, a less suitable image is selected. However, Table I shows that our method often improves the response rather than making it worse, and Table III demonstrates that this approach is most effective in cases where the error of the original model is large (five meters or greater).

[1]https://github.com/cds-mipt/HPointLoc
[2]https://github.com/HRNet/HRNet-Semantic-Segmentation
[3]https://github.com/cvg/Hierarchical-Localization
[4]https://github.com/Nanne/pytorch-NetVlad
[5]https://github.com/QVPR/Patch-NetVLAD

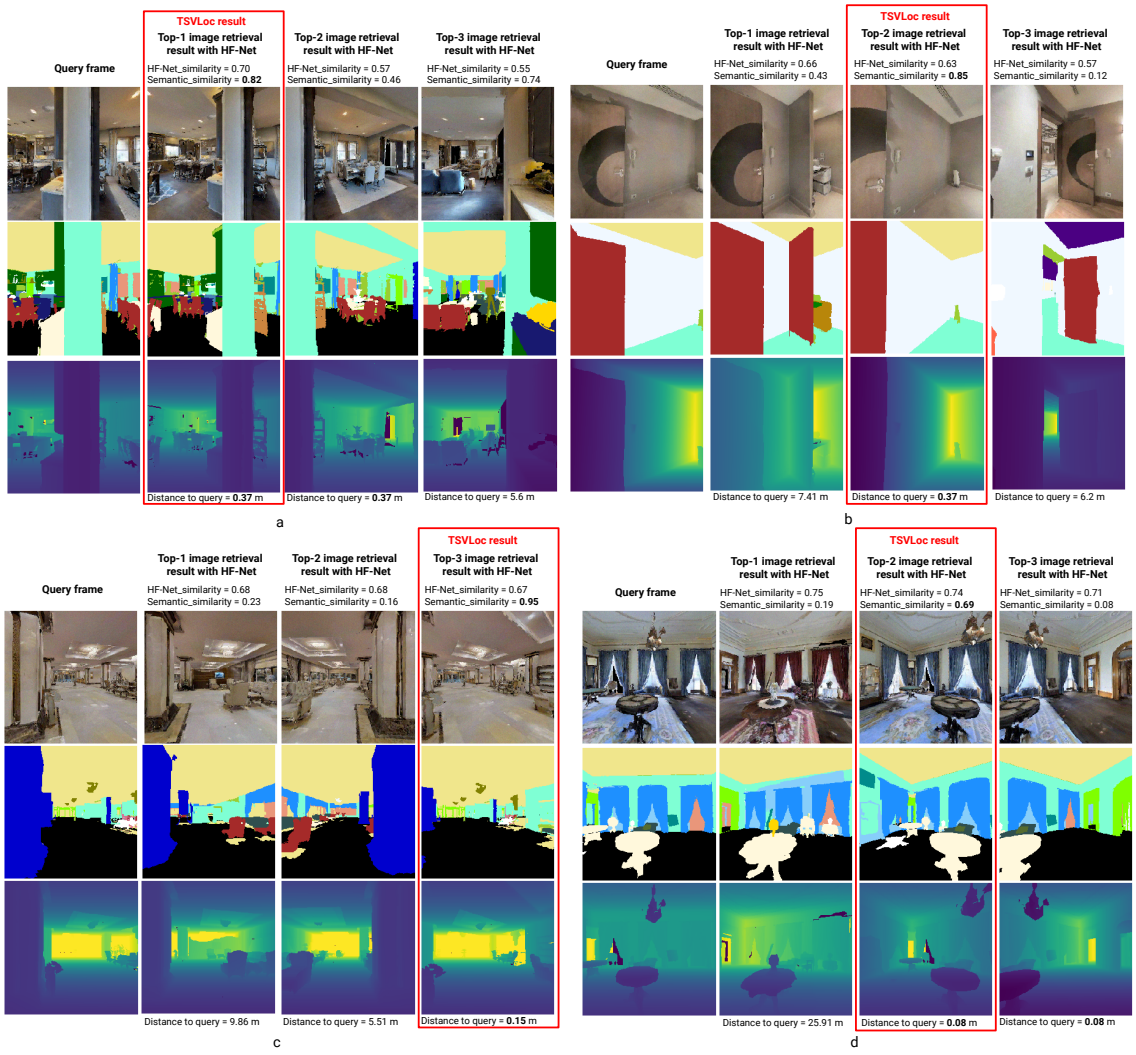

Fig. 4. Examples of successful image retrieval results for proposed TSVLoc method

TABLE III
LOCALIZATION METRICS ON THE SAMPLES OF HPOINTLOC-HARD SUBSET.

Method	R(0.5)	R(1)	R(5)	R(10)
HF-Net	0.000	0.011	0.632	0.780
TSVLoc(H+SEB)	0.085	0.097	0.799	0.897
TSVLoc(H+SEBD)	**0.087**	**0.098**	**0.802**	**0.913**

Fig. 4 shows some examples of successful image retrieval results of our method. Fig. 4 (a) demonstrates the case when both the TSVLoc and the base method give the optimal answer. Other examples demonstrate cases where TSVLoc chooses a better answer than the top-1 result of HF-Net. In the case from Fig. 4 (b,c,d), it can be seen that the HF-Net chose the top-1 image as the most suitable, probably because of the door in a similar perspective, although it is clear that these rooms are different because of the absence of a wall and objects at the corner of the room in the query image. These examples attest to the benefits of using semantic scene encoding. Fig. 5 shows an example for which both the base method and ours offer a far-from-the-optimal-answer; one of the possible ways to solve such cases is using a more complex scene encoding.

C. Ablation Studies

Applying the semantic representation of scenes for the place recognition task, we started with the simplest representation of scenes and gradually added various improvements to it, achieving an increase in localization metrics with TSVLoc: the base variant is a simple enumeration of the types of objects represented in the image and encoding them into one vector (SE); the first improvement to this method is to encode every pair of objects with common boundaries into one entity through an auxiliary vector (SEB); the next improvement involves the center of mass being calculated for each instance of segmentation, and the depth value at the point of the center of mass being encoded into a vector of this object (SEBD). Table IV shows the localization metrics in the case of using only semantic vectors. Here the answer is an image with the

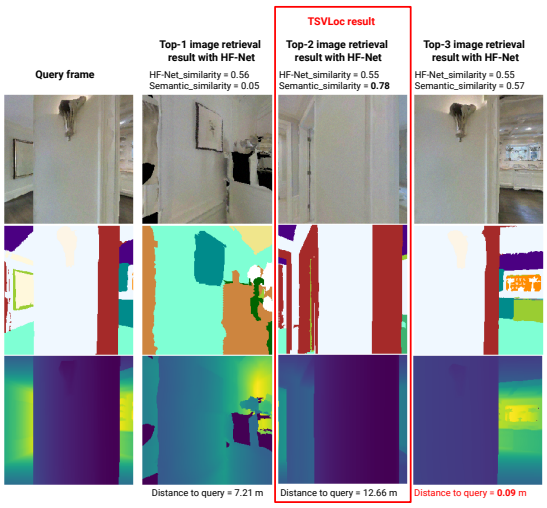

Fig. 5. Example of erroneous result for TSVLoc method

closest semantic vector to the query vector in terms of dot product.

TABLE IV
LOCALIZATION METRICS ON ALL QUERY IMAGES FROM THE HPOINTLOC DATASET USING ONLY SEMANTIC VECTORS

Encoding type	R(0.5)	R(1)	R(5)	R(10)
SE	0.294	0.295	0.644	0.771
SEB	0.357	0.359	0.669	0.821
SEBD	**0.372**	**0.375**	**0.679**	**0.834**

The choice of a localization result with a semantic approach occurs in two stages: in the first stage, images from the database are ranked by their similarity to the query image (using the HF-Net or NetVlad method); in the second stage, topN instances are taken from the ranked images. The first image is considered as a result. During the transition to the next image, if the similarity of the base descriptors decreased by less than γ_1 relative to the current result and the semantic similarity increased by more than γ_2, this image is considered as a result. N, γ_1, and γ_2 are tunable hyperparameters. We attained the best results using this two-step method; besides, we tried to recalculate the similarities of the images s as follows:

$$S = (1 - \alpha)S_{base} + \alpha S_{sem},$$

where S_{base} is the similarity taken from the base method, S_{sem} is the similarity of semantic vectors. After this recalculation, the images were re-ranked and the top-1 image was selected. The best results were achieved with $\alpha = 0.1$. This method showed worse results than the two-stage method but provided a minor increase to the metrics. Table V lists the results of using this approach.

Additionally, we conducted experiments with different dimensions of the semantic vector E_{sem}. Fig. 6 shows the value of the metrics on the HPointLock dataset for the model TSVLoc(H+SEBD). The values reach a plateau at a dimension of 1,000.

TABLE V
LOCALIZATION METRICS ON ALL QUERY IMAGES FROM THE HPOINTLOC DATASET WITH RECALCULATED SIMILARITIES ($\alpha = 0.1$)

Method	R(0.5)	R(1)	R(5)	R(10)
HF-Net	0.890	0.892	0.963	0.976
HF-Net + SEB	0.890	0.892	0.965	0.980
HF-Net + SEBD	**0.891**	**0.893**	**0.968**	**0.981**

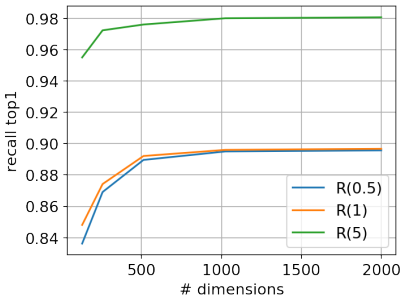

Fig. 6. Recall metric with different thresholds on the HPointLock dataset for the model TSVLoc(H+SEBD) with varied dimensions of the semantic vector E_{sem}

V. CONCLUSION

The key result of this work is that we proposed a novel Two-Stage Vector Symbolic approach (TSVLoc) to construct semantic scene representation based on input semantic and depth maps. Experiments have shown that the TSVLoc method of semantic image retrieval significantly improves previous methods based on the popular neural network models HF-Net, NetVLAD, and Patch-NetVLAD.

Thus, the generation of additional embeddings using Vector Symbolic Architectures based on segmentation and depth maps (SEBD mode of our TSVLoc approach) offers a more accurate solution to the problem of rough global localization. This proves to be promising for the loop detection and first approximation of the camera pose in the methods of simultaneous localization and mapping. The ability to decode the generated semantic embedding by VSA operations into HD vectors from which it was constructed can enhance the interpretability of the selection of the most similar image from the database. It will be the subject of further research.

REFERENCES

[1] S. Hausler, S. Garg, M. Xu, M. Milford, and T. Fischer, "Patch-netvlad: Multi-scale fusion of locally-global descriptors for place recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 14 141–14 152.

[2] D. Kleyko, M. Davies, E. P. Frady, P. Kanerva, S. J. Kent, B. A. Olshausen, E. Osipov, J. M. Rabaey, D. A. Rachkovskij, A. Rahimi, and F. T. Sommer, "Vector symbolic architectures as a computing framework for nanoscale hardware," 2021.

[3] T. A. Plate, "Estimating analogical similarity by dot-products of holographic reduced representations," in *Proceedings of the 6th International Conference on Neural Information Processing Systems*, ser. NIPS'93. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1993, p. 1109–1116.

[4] D. A. Rachkovskij and E. M. Kussul, "Binding and normalization of binary sparse distributed representations by context-dependent thinning," *Neural Computation*, vol. 13, no. 2, pp. 411–452, 2001.

[5] R. Gayler, "Multiplicative binding, representation operators, and analogy," in *Advances in Analogy Research: Integration of Theory and Data from the Cognitive, Computational, and Neural Sciences*, 01 1998, pp. 1–4.

[6] C. Eliasmith, *How to build a brain: A neural architecture for biological cognition*. Oxford University Press, 2013.

[7] E. P. Frady, D. Kleyko, and F. T. Sommer, "A Theory of Sequence Indexing and Working Memory in Recurrent Neural Networks," *Neural Computation*, vol. 30, no. 6, pp. 1449–1513, 06 2018. [Online]. Available: https://doi.org/10.1162/neco_a_01084

[8] P. Kanerva, "Computing with high-dimensional vectors," *IEEE Design Test*, vol. 36, no. 3, pp. 7–14, 2019.

[9] J. K. Guo, D. Van Brackle, N. Lofaso, and M. O. Hofmann, "Vector representation for sub-graph encoding to resolve entities," *Procedia Computer Science*, vol. 95, pp. 327–334, 2016, complex Adaptive Systems Los Angeles, CA November 2-4, 2016. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1877050916325157

[10] E. P. Frady, S. J. Kent, B. A. Olshausen, and F. T. Sommer, "Resonator Networks, 1: An Efficient Solution for Factoring High-Dimensional, Distributed Representations of Data Structures," *Neural Computation*, vol. 32, no. 12, pp. 2311–2331, 12 2020. [Online]. Available: https://doi.org/10.1162/neco_a_01331

[11] E. Osipov, D. Kleyko, and A. Legalov, "Associative synthesis of finite state automata model of a controlled object with hyperdimensional computing," in *IECON 2017 - 43rd Annual Conference of the IEEE Industrial Electronics Society*, 2017, pp. 3276–3281.

[12] D. Gálvez-López and J. D. Tardos, "Bags of binary words for fast place recognition in image sequences," *IEEE Transactions on Robotics*, vol. 28, no. 5, pp. 1188–1197, 2012.

[13] R. Zhao and K. Mao, "Fuzzy bag-of-words model for document representation," *IEEE transactions on fuzzy systems*, vol. 26, no. 2, pp. 794–804, 2017.

[14] R. Arandjelovic and A. Zisserman, "All about vlad," in *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 2013, pp. 1578–1585.

[15] R. Arandjelović, P. Gronat, A. Torii, T. Pajdla, and J. Sivic, "Netvlad: Cnn architecture for weakly supervised place recognition," 2016.

[16] P.-E. Sarlin, C. Cadena, R. Siegwart, and M. Dymczyk, "From coarse to fine: Robust hierarchical localization at large scale," 2019.

[17] J. Revaud, J. Almazan, R. S. de Rezende, and C. R. de Souza, "Learning with average precision: Training image retrieval with a listwise loss," 2019.

[18] P. Neubert, S. Schubert, K. Schlegel, and P. Protzel, "Vector semantic representations as descriptors for visual place recognition," in *Proc. of Robotics: Science and Systems (RSS)*, 2021.

[19] C. Toft, E. Stenborg, L. Hammarstrand, L. Brynte, M. Pollefeys, T. Sattler, and F. Kahl, "Semantic match consistency for long-term visual localization," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 383–399.

[20] J. Johnson, R. Krishna, M. Stark, L.-J. Li, D. A. Shamma, M. S. Bernstein, and L. Fei-Fei, "Image retrieval using scene graphs," in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 3668–3678.

[21] R. Herzig, M. Raboh, G. Chechik, J. Berant, and A. Globerson, "Mapping images to scene graphs with permutation-invariant structured prediction," in *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, ser. NIPS'18. Red Hook, NY, USA: Curran Associates Inc., 2018, p. 7211–7221.

[22] J. Yang, J. Lu, S. Lee, D. Batra, and D. Parikh, "Graph r-cnn for scene graph generation," in *Proceedings of the European Conference on Computer Vision (ECCV)*, September 2018.

[23] M. Qi, W. Li, Z. Yang, Y. Wang, and J. Luo, "Attentive relational networks for mapping images to scene graphs," in *Conference on Computer Vision and Pattern Recognition (CVPR)*, 06 2019, pp. 3952–3961.

[24] H. Dhamo, A. Farshad, I. Laina, N. Navab, G. D. Hager, F. Tombari, and C. Rupprecht, "Semantic image manipulation using scene graphs," in *CVPR*, 2020.

[25] D. Xu, Y. Zhu, C. B. Choy, and L. Fei-Fei, "Scene graph generation by iterative message passing," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 3097–3106.

[26] R. Zellers, M. Yatskar, S. Thomson, and Y. Choi, "Neural motifs: Scene graph parsing with global context," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018, pp. 5831–5840.

[27] Y. Li, W. Ouyang, Z. Bolei, S. Jianping, Z. Chao, and X. Wang, "Factorizable net: An efficient subgraph-based framework for scene graph generation," in *ECCV*, 2018.

[28] W. Wang, R. Wang, S. Shan, and X. Chen, "Exploring context and visual pattern of relationship for scene graph generation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[29] X. Lin, C. Ding, J. Zeng, and D. Tao, "Gps-net: Graph property sensing network for scene graph generation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[30] I. Armeni, Z.-Y. He, J. Gwak, A. R. Zamir, M. Fischer, J. Malik, and S. Savarese, "3d scene graph: A structure for unified semantics, 3d space, and camera," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019.

[31] J. Wald, H. Dhamo, N. Navab, and F. Tombari, "Learning 3d semantic scene graphs from 3d indoor reconstructions," in *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[32] J. Gu, H. Zhao, Z. L. Lin, S. Li, J. Cai, and M. Ling, "Scene graph generation with external knowledge and image reconstruction," *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1969–1978, 2019.

[33] R. Krishna, Y. Zhu, O. Groth, J. Johnson, K. Hata, J. Kravitz, S. Chen, Y. Kalantidis, L.-J. Li, D. A. Shamma, M. Bernstein, and L. Fei-Fei, "Visual genome: Connecting language and vision using crowdsourced dense image annotations," 2016. [Online]. Available: https://arxiv.org/abs/1602.07332

[34] A. Kuznetsova, H. Rom, N. Alldrin, J. Uijlings, I. Krasin, J. Pont-Tuset, S. Kamali, S. Popov, M. Malloci, A. Kolesnikov, T. Duerig, and V. Ferrari, "The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale," *IJCV*, 2020.

[35] M. Ledoux, "The concentration of measure phenomenon," *AMS Surveys and Monographs*, vol. 89, 01 2001.

[36] A. N. Gorban and I. Y. Tyukin, "Blessing of dimensionality: mathematical foundations of the statistical physics of data," *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, vol. 376, no. 2118, p. 20170237, 2018. [Online]. Available: https://royalsocietypublishing.org/doi/abs/10.1098/rsta.2017.0237

[37] G. Montone, J. O'Regan, and A. Terekhov, "Hyper-dimensional computing for a visual question-answering system that is trainable end-to-end," *ArXiv*, vol. abs/1711.10185, 2017.

[38] D. Kleyko, E. Osipov, R. W. Gayler, A. I. Khan, and A. G. Dyer, "Imitation of honey bees' concept learning processes using vector symbolic architectures," *Biologically Inspired Cognitive Architectures*, vol. 14, pp. 57–72, 2015. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2212683X15000456

[39] A. K. Kovalev, A. I. Panov, and E. Osipov, "Hyperdimensional representations in semiotic approach to agi," in *Artificial General Intelligence*, B. Goertzel, A. I. Panov, A. Potapov, and R. Yampolskiy, Eds. Cham: Springer International Publishing, 2020, pp. 231–241.

[40] B. Komer, T. C. Stewart, A. R. Voelker, and C. Eliasmith, "A neural representation of continuous space using fractional binding," in *41st annual meeting of the cognitive science society*. QC: Cognitive Science Society, 2019.

[41] H. Noh, A. Araujo, J. Sim, T. Weyand, and B. Han, "Large-scale image retrieval with attentive deep local features," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 3456–3465.

[42] M. Savva, A. Kadian, O. Maksymets, Y. Zhao, E. Wijmans, B. Jain, J. Straub, J. Liu, V. Koltun, J. Malik, D. Parikh, and D. Batra, "Habitat: A Platform for Embodied AI Research," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019.

[43] A. Chang, A. Dai, T. Funkhouser, M. Halber, M. Niessner, M. Savva, S. Song, A. Zeng, and Y. Zhang, "Matterport3d: Learning from rgb-d data in indoor environments," *arXiv preprint arXiv:1709.06158*, 2017.

[44] W. Maddern, G. Pascoe, C. Linegar, and P. Newman, "1 Year, 1000km: The Oxford RobotCar Dataset," *The International Journal of Robotics Research (IJRR)*, vol. 36, no. 1, pp. 3–15, 2017. [Online]. Available: http://dx.doi.org/10.1177/0278364916679498

[45] G. Neuhold, T. Ollmann, S. Rota Bulo, and P. Kontschieder, "The mapillary vistas dataset for semantic understanding of street scenes," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 4990–4999.