Evaluating the quality of robotic visual-language maps

Matti Pekkanen, Tsvetomila Mihaylova, Francesco Verdoja, and Ville Kyrki

***Abstract*— Visual-language models (VLMs) have recently been introduced in robotic mapping by using the latent representations, *i.e.*, embeddings, of the VLMs to represent the natural language semantics in the map. The main benefit is moving beyond a small set of human-created labels toward open-vocabulary scene understanding. While there is anecdotal evidence that maps built this way support downstream tasks, such as navigation, rigorous analysis of the quality of the maps using these embeddings is lacking. In this paper, we propose a way to analyze the quality of maps created using VLMs by evaluating two critical properties: queryability and consistency. We demonstrate the proposed method by evaluating the maps created by two state-of-the-art methods, VLMaps and OpenScene, using two encoders, LSeg and OpenSeg, using real-world data from the Matterport3D data set. We find that OpenScene outperforms VLMaps with both encoders, and LSeg outperforms OpenSeg with both methods.**

I. Introduction

Mobile robots must understand the geometry and semantics of the environment to accomplish complex tasks. While semantic mapping is an established field, most methods are only able to segment the map into a small pre-selected set of categories. This imposes challenges in real-world settings when a robot operates in an environment that the categories do not fully describe [1], [2].

Visual-Language Models (VLMs), such as CLIP [3], are networks that jointly train a visual and language encoder with large amounts of data to learn a mapping of text and image inputs into a common visual-language latent space. This enables these models to effectively have an *open vocabulary*, meaning visual semantics can be matched to natural language instead of a set of selected symbols. Therefore, the exploitation of VLMs in robotic mapping could allow richer semantics to be represented and to overcome the difficulty of adapting to new environments. We call maps using the embeddings of VLMs as semantics *visual-language maps*.

Complex *queries* have been long possible in semantic maps by inference using object ontologies [4]. However, while semantic labels support only binary comparison as they are enumerations with no notion of distance, visual-language maps can evaluate the query by directly comparing the embedding of the query to the embeddings of the map with a similarity metric. Several successful algorithms have been built on VLMs that can perform mobile robot tasks, such as navigation [5]–[7]. However, the analysis from the robotic mapping perspective is left anecdotal.

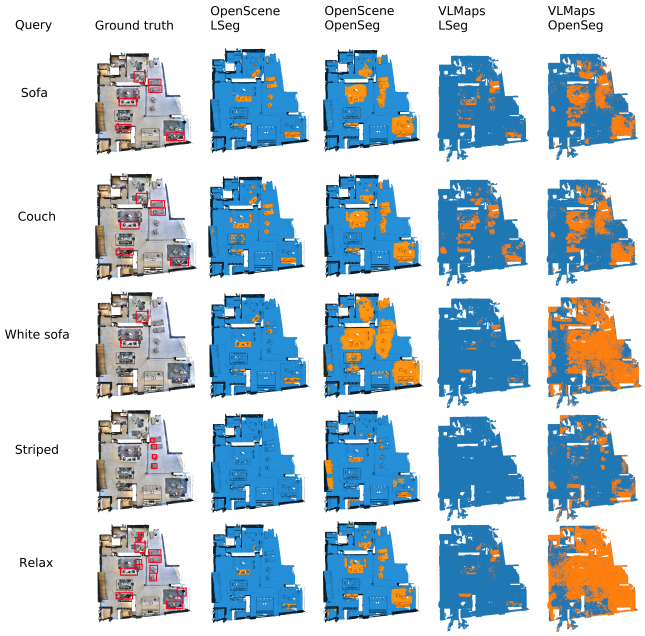


Fig. 1: Maps built from visual-language models have the ability to represent complex semantics, encompassing both the class of objects and their properties. We propose a benchmark for these type of maps and evaluate the quality of different state-of-the-art methods according to the consistency of their representation and their ability to be queried.

In this work, we propose a way to evaluate the quality of visual-language maps. We consider that two aspects of the map capture its quality: queryability and consistency. We evaluate queryability such that the evaluation is not restricted to semantic segmentation of the map, which imposes that the map must be partitioned. Instead, each query is evaluated separately. Furthermore, we evaluate the consistency of the representation, *i.e.*, the embeddings themselves. To demonstrate our method, we evaluate two state-of-the-art methods: VLMaps [5] and OpenScene [8], using the Matterport3D [9] data set.

The main contributions of this paper are:

i) We propose a way to analyze the quality of visual-language maps by evaluating the queryability and consistency of the representation
ii) We demonstrate our analysis method by evaluating two state-of-the-art methods

This work was supported by Business Finland (decision 9249/31/2021), the Research Council of Finland (decision 354909), Wallenberg AI, Autonomous Systems and Software Program, WASP and Saab AB. We gratefully acknowledge the support of NVIDIA Corporation with the donation of the Titan Xp GPU used for this research.

M. Pekkanen, T. Mihaylova, F. Verdoja and V. Kyrki are with School of Electrical Engineering, Aalto University, Espoo, Finland. {firstname.lastname}@aalto.fi

II. Related work

Creating maps using embeddings of VLMs, especially CLIP [3], has recently gained attention, with the proposal of methods such as VLMaps [5], OpenScene [8], NLMaps [10], and Uni-Fusion [11].

In this paper, we focus on VLMaps and OpenScene. Both create a dense grid map, where each cell is associated with a single VLM embedding. The embeddings are created from RGBD images with VLM-based semantic segmentation encoder, such as LSeg [12] or OpenSeg [13], and back-projected to the 3D map. Each cell is represented as the mean of the embeddings projected to the cell. In the original work, the 3D map in VLMaps is projected to 2D by averaging the embeddings in the z-axis. Additionally, OpenScene learns an encoder that directly produces CLIP embeddings from the 3D point cloud. Combining these two approaches, they propose joint 2D-3D "ensemble" features as the final representations in the map.

In VLMaps, as with other zero-shot navigation methods [6], [7], the map quality is not directly assessed, but instead, they use the success of the downstream navigation task as the metric, which does not fully evaluate the capabilities of the representation. Because the metrics used in evaluating visual-language maps in each prior work are different, comparing the quality of their maps is non-trivial, which is the problem we address in this paper.

III. Problem statement

We aim to evaluate the quality of visual-language maps. We only consider methods where the estimation of geometry and semantics are disjoint, so we concentrate on evaluating the semantic quality of the maps. We focus on two aspects of the map that we believe capture its semantic quality: queryability and consistency.

We must evaluate *queryability*, as it is the primary way to retrieve information from the representation and, therefore, acts as the measure of two things: first, indirectly, the ability of the representation to contain relevant information, and second, the accessibility of that information. It also must address *consistency*, as consistent performance within and between measuring runs, times, environments, and sensors is desirable.

IV. Methods

A. Evaluating queryability

1) Voxel-based queryability: The first method evaluates the overall matching between the query results compared to the ground truth in a binary classification setting. This property cannot be evaluated using standard multi-class classification, as the queries do not form a partition of the map. Each query produces a query result, a binary mask, over the whole map, and the same voxel might be a match of multiple queries.

The method consists of the following steps:

i) A map m is created from each sequence from the data set, forming the set $\mathcal{M}$.
ii) Each map $m \in \mathcal{M}$ is queried with each query q in the set of queries $\mathcal{Q}$. The query result of a single query is a binary segmentation mask of the map, $\hat{\mathbf{y}}_q = \{\hat{y}_1, \ldots, \hat{y}_N\}$, consisting of N voxel masks $\hat{y} \in \{true, false\}$. Combined, the query results form the set of voxel-based predictions $\hat{Y}_V = \{\hat{\mathbf{y}}_q \forall q \in \mathcal{Q}\}$.
iii) For each query, the true mask $\mathbf{y}_q$ is created from a ground truth map m_g by selecting the voxels answering query q. The true masks form a set of true labels $Y_V = \{\mathbf{y}_q \forall q \in \mathcal{Q}\}$.
iv) The binary classification metrics are calculated between Y_V and $\hat{Y}_V$.

The metrics used to evaluate the classification tasks are F1-score, Intersection over Union (IoU), precision, and recall. As the regions of interest are relatively small with most queries, true negative predictions dominate the predictions. Therefore, accuracy cannot be used as a metric, as it is calculated using true negative predictions. The other metrics are not affected by true negative predictions, so we use them as the metrics for binary classification tasks in this work.

2) Instance-based queryability: The second method evaluates the capability of the map to detect and retrieve all matching objects for the query. Each instance is predicted to match or not match the query; therefore, this measures the coverage of matches within an object rather than over the whole map.

The method consists of the following steps:

i) Given $\mathcal{M}$, $\mathcal{Q}$, and ground truth instance segmentation $\mathcal{I}$, each map $m \in \mathcal{M}$ is queried with the each query $q \in \mathcal{Q}$, each query yielding binary predicted mask $\hat{\mathbf{y}}_q$.
ii) For each object instance $i \in \mathcal{I}$, the corresponding voxels $\hat{\mathbf{i}}_{i,q} \subset \hat{\mathbf{y}}_q$ are selected from the predicted map.
iii) If the majority of the voxels in $\hat{\mathbf{i}}$ are *true*, the prediction $\hat{y}_{i,q}$ is *true*, otherwise *false*. The predictions for each instance for each query combined form the set of predictions $\hat{Y}_I = \{\hat{y}_{i,q} \forall i \in \mathcal{I}; q \in \mathcal{Q}\}$.
iv) Similar to the previous method, the true map m_q is created. If the instance i on the true map answers the query q, the true label $y_{i,q} = true$, otherwise $false$. The true labels for each instance for each query form the set of all true labels denoted Y_I.
v) The binary classification metrics are calculated between Y_I and $\hat{Y}_I$.

From Y_I and $\hat{Y}_I$, the same metrics presented for the voxel-based queryability are computed for each instance.

B. Evaluating consistency

We further subdivide the consistency of the map into two distinct aspects, intra-map consistency, and inter-map consistency, and propose a method for evaluating each.

1) Intra-map consistency: Intra-map consistency captures the similarity of embeddings across the voxels within a map, sharing semantic meaning. The hypothesis is that embeddings sharing semantic meaning are clustered together in the latent space to allow the separability of different concepts. While each object instance is in some sense unique,

a consistent set of embeddings represents an abstract base class to which the instances belong.

The method consists of the following steps:

i) Given a map m, and semantic label l, $\mathcal{E}_l$ is the set of embeddings corresponding to voxels in the map m where the ground truth semantic label is l, and $\mathcal{E}_m$ is the set of all embeddings in the map m. We form a set of tuples $\mathcal{T} = \{(\mathcal{E}_l, \mathcal{E}_m) \forall m \in \mathcal{M}; l \in \mathcal{L}\}$, where $\mathcal{L}$ is a closed-set semantic label vocabolary. For computational efficiency and to no considerable change in results, in practice, the sets $\mathcal{E}_l$ and $\mathcal{E}_m$ are subsampled. We use a subsampling ratio of 0.1 in this work.

ii) The average absolute deviation d_l^m is calculated for $\mathcal{E}_l$, and d_m^m for $\mathcal{E}_m$, for each tuple $t \in \mathcal{T}$. We use the mean of the embeddings as the central point and cosine distance as the deviation metric, which is consistent with the use of cosine distance loss for CLIP.

iii) The intra-map consistency ratio $c_l^m = \frac{d_l^m}{d_m^m}$ is computed for each tuple $t \in \mathcal{T}$, with label l in map m. This ratio represents the distinguishability of the class compared to the map average.

2) Inter-map consistency: Inter-map consistency captures the similarity of voxels with the same semantic label across different maps. The hypothesis is that embeddings within the same label are closer to each other across maps than to those with different labels with respect to a distance metric. This would imply that the objects retain their distinctiveness across maps, and therefore, the system generalizes better across different environments.

The method consists of the following steps:

i) Given the set of tuples $\mathcal{T}$, constructed according to Section IV-B.1.

ii) For all pairs of tuples $((\mathcal{E}_{l,1}, \mathcal{E}_{m,1}), (\mathcal{E}_{l,2}, \mathcal{E}_{m,2}))$, parametric Wasserstein 2-distance d_w is calculated between $\mathcal{E}_{l,1}$ and $\mathcal{E}_{l,2}$.

The Wasserstein p-distance is computationally heavy for large sets of high-dimensional embeddings. For this reason, given an n-dimensional embedding $e \in \mathbb{R}^n$, we approximate the set of embeddings $\mathcal{E}$ with an n-dimensional normal distribution $\mathcal{N}(\mu, P)$. This allows us to have a closed-form solution for the Wasserstein 2-distance.

V. Experiments

The two main questions the benchmark aims to answer with the experiments are:

1) How queryable state-of-the-art visual-language maps are?
2) How consistent are their visual-language embeddings within and across maps?

To answer these questions, we evaluated two state-of-the-art methods, VLMaps [5] and OpenScene [8], in a series of experiments.

A. Data set

In the evaluation, the same sequences from the same data set, Matterport3D [9], and the same set of 42 labels $\mathcal{L}$ that are used in the original work of VLMaps. The names of the labels form the query set $\mathcal{Q}$, used in the following experiments.

Based on the ground truth semantics, instances used as ground truth were segmented using a region growing algorithm [14]. Each voxel is initialized as a seed cluster; then, region growing steps are performed, where the labels of all neighboring clusters are compared. If the labels are the same, the clusters are joined. Otherwise, they are not. This step is iterated until no more clusters can be joined or a maximum iteration limit is reached.

B. Parametrization of the methods

The VLMaps map was created using the new 3D mapping method available at the repository of the original authors [15], where the central idea is the same, except the problems of aggregating the embeddings along the z-axis are avoided. All of the parameters were the default used by the original authors.

When queried, VLMaps creates a list of queries by setting the original query and string "other" into a list of phrases. Then, the most similar label for each voxel is selected for the set of queries. The binary mask is created by binary comparison between the acquired labels and the query. The binary mask is then dilated to encompass whole object instances.

OpenScene maps were created using the proposed parameters from the original paper. We use the proposed 2D-3D ensemble features. OpenScene does not provide a way to query open-vocabulary binary masks. Because we use the set of labels as the queries, the binary query mask is created by comparing the equality of the voxel labels and the query.

Both methods used the same pre-trained LSeg and OpenSeg encoder provided by the original authors. Both encoders are based on the CLIP backbone; LSeg uses the CLIP-ViT-B/32 backbone, whereas OpenSeg uses the CLIP-ViT-L/14.

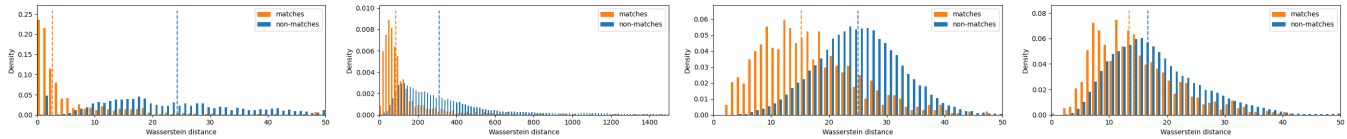
C. Results

1) Queryability: The results of the queryability benchmark are presented in Table I, where the F1-score, precision, recall, and IoU of the method and encoder combinations are presented for the voxel- and instance-based tests. Overall, OpenScene performs better than VLMaps with both encoders, based on the higher F1-scores and IoUs. The higher recall showcased by VLMaps with LSeg is the result of the post-processing of the queries, where the query results are dilated to encompass whole objects. The instance-based results mostly align with the voxel-based results, except the performance is slightly decreased in all metrics.

Therefore, these results show that both the mapping method and choice of encoder matter in the map creation process. They also indicate that the 3D structure of the environment (included by OpenScene) contains information that can be leveraged in addition to the purely image-based creation of embeddings.

TABLE I: Results of the voxel- and instance-based queryability tests.

Method	Encoder	Voxel-based F1	Voxel-based Precision	Voxel-based Recall	Voxel-based IoU	Instance-based F1	Instance-based Precision	Instance-based Recall	Instance-based IoU
OpenScene	LSeg	**0.623**	**0.617**	0.630	**0.457**	0.258	0.252	0.264	0.149
OpenScene	OpenSeg	0.580	0.575	0.587	0.412	**0.261**	**0.255**	0.267	**0.150**
VLMaps	LSeg	0.498	0.401	**0.661**	0.332	0.257	0.205	**0.347**	0.147
VLMaps	OpenSeg	0.393	0.287	0.625	0.246	0.202	0.144	0.342	0.112



(a) OpenScene with LSeg (b) OpenScene with OpenSeg (c) VLMaps with LSeg (d) VLMaps with OpenSeg

Fig. 2: The histograms present the distribution of Wasserstein distances of matching and non-matching labels in orange and blue, respectively. The medians of the distributions are presented with a dashed line. The means were biased by large outliers in OpenScene, preventing their use in visualization.

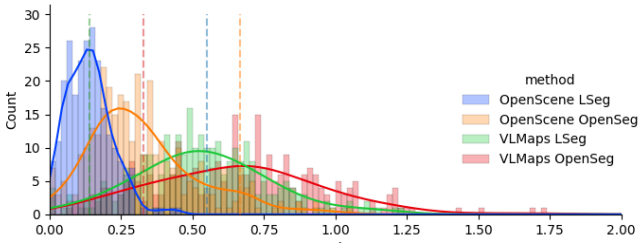


Fig. 3: The intra-map consistency ratios of the methods as histograms, with a kernel density estimation, which is smoothed with a Gaussian kernel. The mean of the distribution is depicted with a vertical line of corresponding color.

Additionally, we qualitatively demonstrate the capability of the benchmark to process open-vocabulary queries. We use the VLMaps' binary masking method without prompt engineering or dilation as the comparison method. From the results presented in Figure 1, it can be seen that OpenScene query results are less noisy, indicating that the features are more distinctive. LSeg yields the objects more accurately, but the more abstract results, such as "striped" or "relax", produce no results. OpenScene is more sensitive, finding solutions to abstract queries but yielding many false positive results.

2) Consistency:

a) The intra-map consistency: The results are presented in Figure 3, in which the intra-map consistency ratios of the methods are illustrated with histograms. Additionally, the means of data are shown with dashed lines. Notably, the ordering of the methods according to their instra-map consistency is the same as in the queryability experiment presented in Section V-C.1. OpenScene has better consistency between the labels with each encoder, which is indicated by the lower means of the distributions. This indicates that OpenScene can cluster the embeddings of labels better using the ensemble features. VLMaps with OpenSeg has a long tail that continues beyond the figure, with several outliers.

b) The inter-map consistency: In Figure 2, the Wasserstein distances between matching, *i.e.*, the label is the same, and non-matching sets of embeddings across all maps are shown. The medians of the data are shown with dashed lines. The ratio between the medians of matching and non-matching labels suggests that OpenScene separates the labels better than VLMaps, and LSeg separates them better than OpenSeg. The ratio of medians of OpenScene with LSeg is 9.21, OpenScene with OpenSeg is 3.67, VLMaps with LSeg 1.65, and VLMaps with OpenSeg 1.24, which is once again the exact ordering previously observed in Sections V-C.1 and V-C.2.a.

While the shapes of distributions are similar, OpenScene has much larger distances on average, even with matching labels. This results from the fact that the distilled features of OpenSeg are of a larger magnitude than the image features of both VLMaps and OpenScene image features. For example, the average norm of all OpenScene with OpenSeg embeddings in the map of the first sequence is 13.68 times larger than the average norm of all VLMaps with OpenSeg embeddings in the same map, and the maximum being 57.97 times larger than the maximum VLMaps feature.

VI. Conclusion

In this work, we present a method for evaluating the quality of visual-language maps. To demonstrate our method, we evaluate two state-of-the-art methods, VLMaps and OpenScene, using two visual-language encoders, LSeg and OpenSeg. We find that OpenScene outperforms VLMaps with both encoders, and LSeg outperforms OpenSeg with both methods.

While we were limited in our quantitative evaluation of closed vocabulary, all the proposed metrics can be extended to open vocabulary. A data set providing ground truth open-vocabulary semantics is direly needed. This would allow extending the evaluation to address the breadth of the vocabulary adequately. With the current open-vocabulary methods evaluated using zero-shot but closed-vocabulary context, the real progress in this aspect is currently unknown. This opens interesting research possibilities to capture the true promise of the VLMs.

REFERENCES

[1] A. Bendale and T. E. Boult, "Towards Open Set Deep Networks," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Las Vegas, NV, USA: IEEE, Jun. 2016, pp. 1563–1572.

[2] C. Geng, S.-J. Huang, and S. Chen, "Recent Advances in Open Set Recognition: A Survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 43, no. 10, pp. 3614–3631, Oct. 2021.

[3] A. Radford *et al.*, "Learning transferable visual models from natural language supervision," in *International Conference on Machine Learning*. PMLR, 2021, pp. 8748–8763.

[4] M. Tenorth and M. Beetz, "KNOWROB - knowledge processing for autonomous personal robots," in *2009 IEEE/RSJ International Conference on Intelligent Robots and Systems*. St. Louis, MO, USA: IEEE, oct 2009, pp. 4261–4266.

[5] C. Huang, O. Mees, A. Zeng, and W. Burgard, "Visual Language Maps for Robot Navigation," in *2023 IEEE International Conference on Robotics and Automation (ICRA)*. London, United Kingdom: IEEE, may 2023, pp. 10 608–10 615.

[6] S. Y. Gadre, M. Wortsman, G. Ilharco, L. Schmidt, and S. Song, "Cows on pasture: Baselines and benchmarks for language-driven zero-shot object navigation," in *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Vancouver, Canada, jun 2023, pp. 23 171–23 181.

[7] D. Shah, B. Osiński, S. Levine *et al.*, "Lm-nav: Robotic navigation with large pre-trained models of language, vision, and action," in *Conference on Robot Learning*. Atlanta, GA, USA: PMLR, nov 2023, pp. 492–504.

[8] S. Peng *et al.*, "Openscene: 3d scene understanding with open vocabularies," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Vancouver, Canada, jun 2023, pp. 815–824.

[9] A. Chang *et al.*, "Matterport3D: Learning from RGB-D Data in Indoor Environments," Sep. 2017, arXiv:1709.06158 [cs].

[10] B. Chen *et al.*, "Open-vocabulary queryable scene representations for real world planning," in *2023 IEEE International Conference on Robotics and Automation (ICRA)*. London, UK: IEEE, may 2023, pp. 11 509–11 522.

[11] Y. Yuan and A. Nüchter, "Uni-fusion: Universal continuous mapping," *IEEE Transactions on Robotics*, vol. 40, pp. 1373–1392, jan 2024.

[12] B. Li, K. Q. Weinberger, S. Belongie, V. Koltun, and R. Ranftl, "Language-driven Semantic Segmentation," Apr. 2022, arXiv:2201.03546 [cs].

[13] G. Ghiasi, X. Gu, Y. Cui, and T.-Y. Lin, "Scaling open-vocabulary image segmentation with image-level labels," in *European Conference on Computer Vision*. Tel Aviv, Israel: Springer, oct 2022, pp. 540–557.

[14] R. Adams and L. Bischof, "Seeded region growing," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 16, no. 6, pp. 641–647, Jun. 1994.

[15] Chenguang Huang and Oier Mees and Andy Zeng and Wolfram Burgard. Vlmaps. [Online]. Available: https://github.com/vlmaps/vlmaps