Revisiting Object-Level Uncertainties for Robust Visual Place Recognition

Alex Junho Lee and Dong jin Hyun*

***Abstract*— Visual place recognition has been challenging and crucial in real-world applications such as autonomous navigation and vision-based robot missions. The introduction of foundation models has greatly enhanced the accuracy of vision-based algorithms and visual place recognition, expanding to the methods that utilize semantic information from images. In this workshop paper, we revisit the features of semantic classification in the visual place recognition process to discuss how to deal with outcomes of semantic segmentation during visual place recognition. By showing that the semantic labels are not uniformly distributed, we propose to handle the uncertainty of semantic classes as a bivariate distribution that depends on the class and the assigned localization clusters instead of commonly used class-level confidences. Utilizing a powerful foundation model capable of language-image similarity evaluation, we evaluate and show the distributions of semantic class activations in the public datasets.**

I. Introduction

Visual Place Recognition (VPR) is essential for vision-based robotic systems and is widely applied in the industry for autonomous navigation, mission planning, map building, digital twin, and augmented reality. While VPR has made significant progress, achieving the ability to work flawlessly over real-world variances remains a challenge. With the introduction of foundation models [1] [2] [3] in computer vision, the accuracy and potential of vision-based algorithms have been largely boosted, and also VPR benefited from the generalization ability of them [4]. Because foundation models were mostly trained by the gigantic size of training data and an efficient representation of them, they show a high-end generalization ability. Technically, VPR is a procedure of retrieving the closest image of query image from database images, which can be reformulated as a N-class classification problem when the database is size of N. A key factor of VPR is building a reliable representation of images to properly encode scene information into un-varying place representation while neglecting the non-place specific and non-distinctive variables.

This image-to-feature-transformation can be interpreted in two stages: image encoding and transformation for localization purposes. Even if the image encoding perfectly represents every object and context in the image, directly matching all the descriptors will not always result in correct matches due to the differences between the database and query. This difference derives from various origins in the real world, such as illumination, seasonal changes, disappearing, and appearing objects over time. Therefore, a proper method of transforming the encoded images into localization descriptors is required to achieve robust localization. Researchers have covered various strategies both on image encoding and clustering [5] [6] [7]. NetVLAD [7] has been widely used since its release. Recently, approaches have been suggested to utilize additional information, such as semantics [8] [9]. Semantic segmentation is one of the most efficient strategies for transforming image descriptors into localization descriptors because variable components are usually produced by the changing objects in the scene (e.g., parked cars in the parking lot and pedestrians on the road). However, the uncertainties of the semantic labels have not been extensively covered. In the conventional approaches, the uncertainty is handled by ignoring some of the "dynamic" labels, such as pedestrians or vehicles. However, the distribution of those dynamic labels is not completely uniform. Still, it is rather a conditional probability that relies on the characteristics of the places, as shown in Fig. 1. For example, a class [car] will be highly variable in the parking lot but will not be seen indoors if it is not a car exhibition. In this workshop paper, we concentrate on the conditional probability of semantic classes appearing in different places and further suggest how to handle the uncertainty of each semantic class using the suggested similarity score and intra-cluster variance.

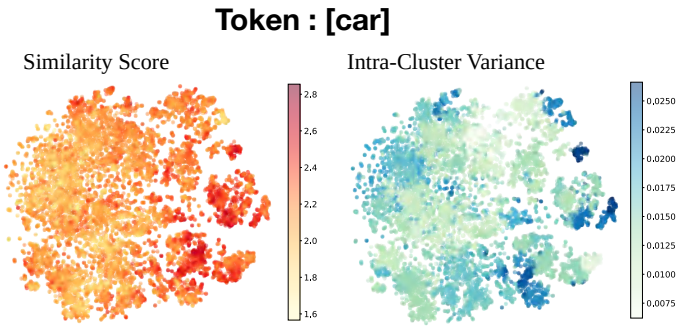

Fig. 1. Activation distribution of the token [car], visualized using t-SNE on VLAD [5] descriptors from the pitts30k dataset's training sequence. Here, a higher similarity score corresponds to a smaller cosine distance between the CLIP image encoder's image embeddings and the CLIP text encoder's text embeddings. As displayed on the left, token [car] activation concentrates primarily within a few clusters. However, as indicated from the right, the variance within their clusters varies, suggesting that activation is meaningful in specific clusters but not others.

*Corresponding author: Dongjin Hyun

All authors are with the Robotic Lab, Research and Development Division, Hyundai Motor Company, Uiwang 16082, Republic of Korea. {alexjunholee, mecjin}@gmail.com

II. Experiments

To calculate the similarity score and intra-cluster variance of suggested semantic classes, we first need to acquire the appearance rate of each text label upon places. For this, we can use datasets such as Cityscapes [10]; however, we can also benefit from the language-image model such as Contrastive Language-Image Pre-Training (CLIP) [3]. As CLIP is capable of calculating the similarity between suggested text labels and images, we utilize the module to evaluate the similarity between suggested text classes and the image. Using this module makes our method applicable to datasets without semantic labels.

We first extract text embedding from the selected class labels from Cityscapes and transform them into text embedding using a CLIP text encoder. Also, the images from the database are transformed into image embedding using an image encoder from CLIP. Then, we calculate the similarity between the descriptors to show the relevance between the suggested keyword class and database image. An image containing contexts related to the suggested test label will show a higher similarity.

With obtained similarity values, we aim to discover the similarity distribution upon the localization vectors. Therefore, we build localization descriptors based on the image embeddings and NetVLAD pipeline, using the CLIP image encoder as the encoding layer and NetVLAD as the pooling layer. After all, we clusterize the training database into 64 clusters using the K-means algorithm as in Fig. 2.

Assuming the database is acquired sufficiently multiple times for places in the training set, we can calculate the mean and variance as in Fig. 3. To visualize the similarity distribution in the vector space, we plotted the similarity over the t-SNE plot as in Fig. 4 and Fig. 5. As observed from the figures, the activation occurs in different clusters depending on the class labels, and the intra-cluster variance also differs. From the higher similarity, we can assume that the class labels will likely be present in the clusters and that the class is relevant to such clusters. However, if the intra-cluster variance is high, the class is often not informative. Therefore, we can trust the class relevance only when the similarity is high and the variance is low.

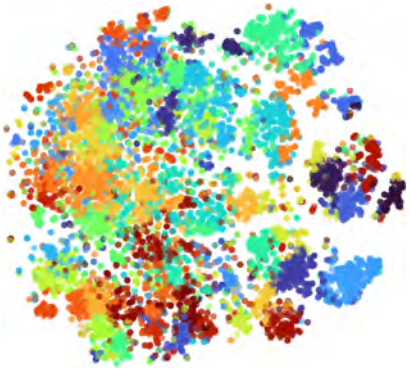

Fig. 2. K-means Clusterization results on VLAD descriptors from the pitts30k dataset's training sequence. VLAD descriptors that are assigned to different clusters are colored in different colors.

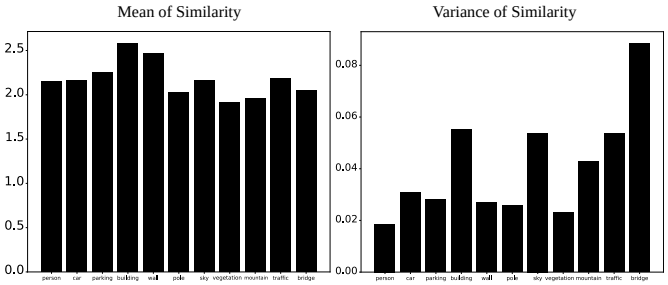

Fig. 3. (left) mean values of similarity and (right) variances of similarity from selected text classes on the pitts30k dataset. In contrast to our intuitions, the higher variance means the text class is more place-specific because the higher value results from the clearly distinguished distribution of places.

III. Conclusion

In this workshop paper, we suggest modeling the reliability of semantic classes for VPR not by class uncertainty but by conditional probability distribution depending on the places and classes. We have shown that the distribution is nonuniform, and classes could provide additional information in some circumstances, even if it varies in most places. In future works, we expect to enhance the performance of VPR by applying adaptation based on this suggested conditional probability.

References

[1] M. Oquab, T. Darcet, T. Moutakanni, H. Vo, M. Szafraniec, V. Khalidov, P. Fernandez, D. Haziza, F. Massa, A. El-Nouby, *et al.*, "Dinov2: Learning robust visual features without supervision," *arXiv preprint arXiv:2304.07193*, 2023.

[2] M. Caron, H. Touvron, I. Misra, H. Jégou, J. Mairal, P. Bojanowski, and A. Joulin, "Emerging properties in self-supervised vision transformers," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 9650–9660.

[3] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, *et al.*, "Learning transferable visual models from natural language supervision," in *International conference on machine learning*. PMLR, 2021, pp. 8748–8763.

[4] N. Keetha, A. Mishra, J. Karhade, K. M. Jatavallabhula, S. Scherer, M. Krishna, and S. Garg, "Anyloc: Towards universal visual place recognition," *IEEE Robotics and Automation Letters*, 2023.

[5] H. Jégou, M. Douze, C. Schmid, and P. Pérez, "Aggregating local descriptors into a compact image representation," in *2010 IEEE computer society conference on computer vision and pattern recognition*. IEEE, 2010, pp. 3304–3311.

[6] F. Radenović, G. Tolias, and O. Chum, "Fine-tuning cnn image retrieval with no human annotation," *IEEE transactions on pattern analysis and machine intelligence*, vol. 41, no. 7, pp. 1655–1668, 2018.

[7] R. Arandjelović, P. Gronat, A. Torii, T. Pajdla, and J. Sivic, "NetVLAD: CNN architecture for weakly supervised place recognition," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2016.

[8] V. Paolicelli, A. Tavera, C. Masone, G. Berton, and B. Caputo, "Learning semantics for visual place recognition through multi-scale attention," in *International Conference on Image Analysis and Processing*. Springer, 2022, pp. 454–466.

[9] G. Peng, Y. Yue, J. Zhang, Z. Wu, X. Tang, and D. Wang, "Semantic reinforced attention learning for visual place recognition," in *2021 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2021, pp. 13 415–13 422.

[10] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele, "The cityscapes dataset for semantic urban scene understanding," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 3213–3223.

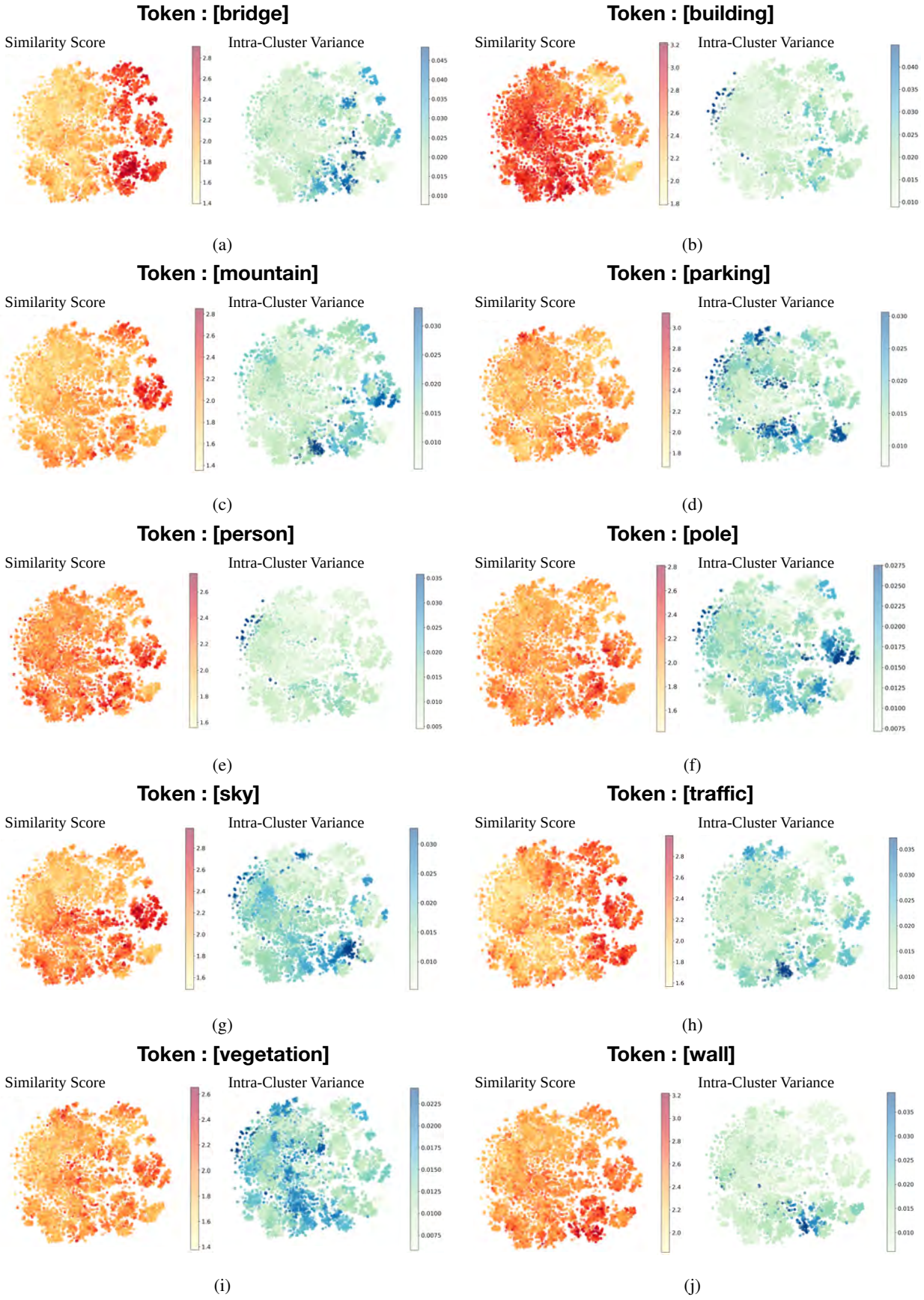

Fig. 4. Activation distribution of selected classes text from CityScapes [10] class definitions, visualized using t-SNE on VLAD descriptors from the pitts30k dataset's training sequence. Depending on the classes, we may observe the distributions as expectations, as classes like "traffic" and "bridge" are cluster-specific, and classes like "building" are non-cluster-specific.

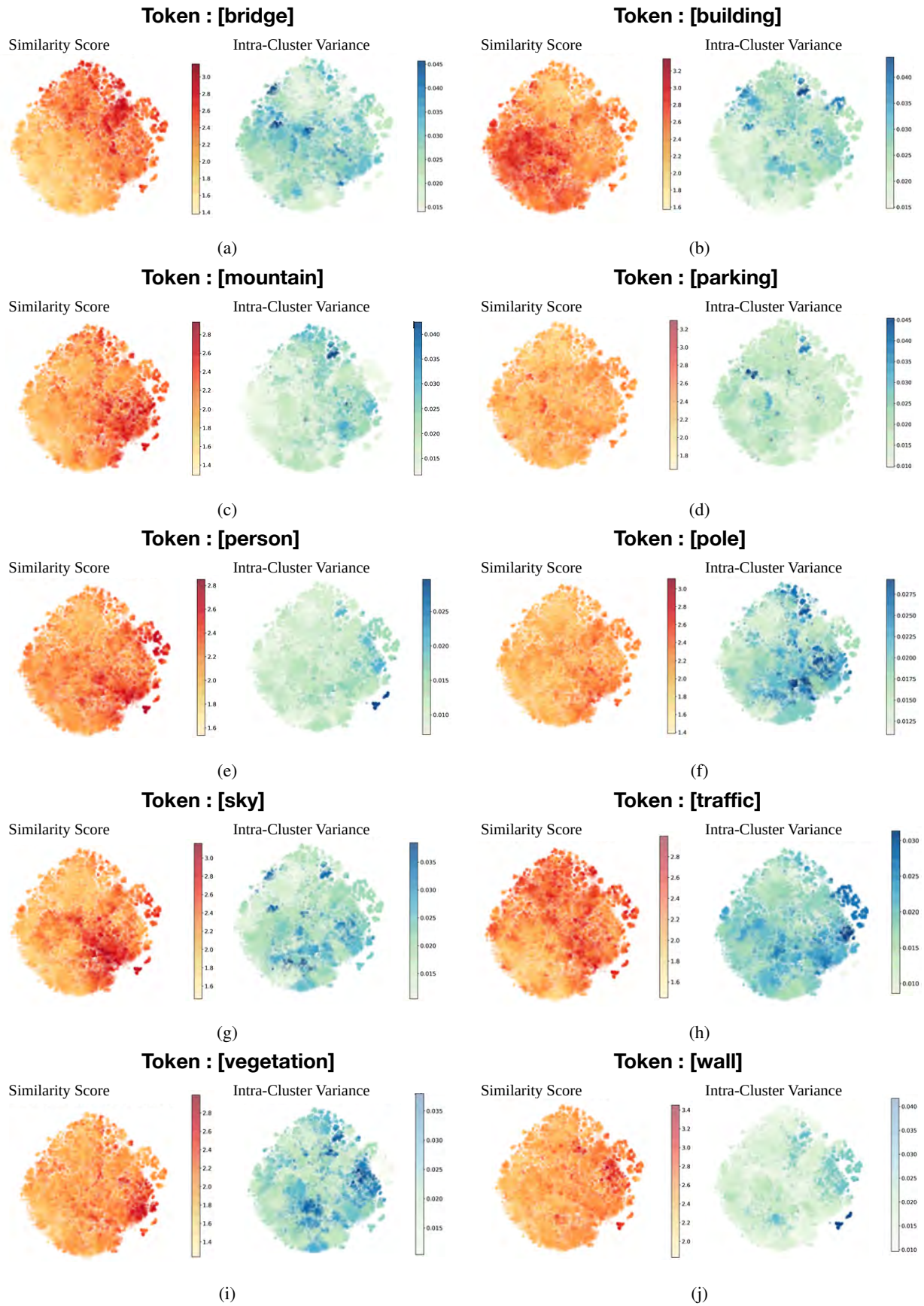

Fig. 5. Activation distribution of selected classes text from CityScapes [10] class definitions, visualized using t-SNE on VLAD descriptors from the pitts250k dataset's training sequence.