Journal of Korea Robotics Society (2026) 21(2):158-163
https://doi.org/10.7746/jkros.2026.21.2.158 ISSN: 1975-6291 / eISSN: 2287-3961

언어 기반 네비게이션에서의 불확실성 재고

Revisiting Uncertainty in Language-driven Navigation

이 준 호[1] · 현 동 진†
Alex Junho Lee[1], Dong Jin Hyun†

Abstract: Language-driven navigation is a key challenge for autonomous robots, requiring the integration of natural language understanding with robust perception and localization. Existing approaches in visual place recognition and navigation mainly rely on similarity-based feature matching, while variance estimation or explicit uncertainty modeling has rarely been a fundamental component. Although a number of prior studies have investigated uncertainty in tasks such as object detection and target search, these efforts remain limited. In this paper, we revisit uncertainty in language-driven navigation and argue that it should be addressed at the distributional level rather than reduced to point estimates such as similarity search. Leveraging a large language model capable of semantic reasoning, we analyze how non-uniform distributions of semantic features and class activations affect navigation performance on benchmark datasets. The results show that conventional similarity-based measures are insufficient to capture the variability inherent in semantic maps, underscoring the necessity of redefining uncertainty modeling as a foundation for reliable perception, robust decision-making, and deployment of language-driven robotic navigation.

Keywords: Navigation, Uncertainty Estimation, Localization, Perception

1. 서론 및 배경

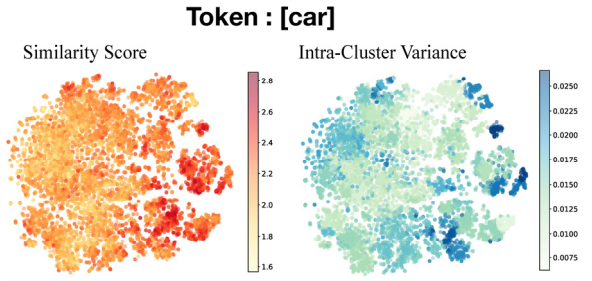
언어 기반 네비게이션(Language-driven Navigation)은 자율주행, 로봇 내비게이션, 증강현실, 디지털 트윈 등 다양한 응용에서 핵심 기능인 장소 인식(place recognition)과 의미적 환경 이해를 동시에 요구한다. 최근 foundation model의 도입[1-3]은 단순 외양 기반 유사도 비교를 넘어 의미적 단서에 근거한 장소 해석을 가능하게 하였으며, 그 결과 공간 정보 정합은 semantic grounding을 포함하는 문제로 재구성되었다. 특히 vision-language 기반 모델은 이미지에 포함된 텍스트 기반 의미 단서(예: car, road, elevator, corridor)를 효과적으로 포착할 수 있어, 단순 유사도 매칭 문제를 공간의 의미적 이해 문제로 확장시킨다.

시각 정보를 활용하는 시각적 장소 인식(Visual Place Recognition, VPR)은 로봇의 위치를 추정하여 카메라 기반 공간 로봇 서비스를 가능하게 하는 기본 구성 요소이다. VPR은 질의 영상(query image)과 데이터베이스(database)에 저장된 장소 표현(place representation)을 비교하여 동일 장소 여부를 판별하는 절차로 정의되며, 다중 분류(N-class classification) 문제로도 해석될 수 있다. 그러나 실제 환경에서는 조명 조건, 날씨 변화, 계절적 요인, 보행자와 차량과 같은 동적 객체 등 다양한 요인으로 인해 동일 장소도 다르게 관측될 수 있어, 강건한 VPR은 여전히 해결해야 할 도전 과제로 남아 있다.

이러한 문제를 극복하기 위해 다양한 이미지 인코딩 및 클러스터링 기법이 제안되었으며, 대표적으로 VLAD 계열 알고리즘이 장소 인식을 위한 전역 표현을 학습하는 데 널리 사용되어 왔다[4,5]. 또한 최근에는 시맨틱 정보를 결합하는 접근이 활발히 시도되고 있다[5-7]. 시맨틱 분할은 장면 내에 존재하는 객체를 클래스 단위로 구분함으로써, 변동성이 큰 동적 객체의 영향을 줄이는 데 효과적이다. 예를 들어, 주차장의 차량이나 도로 위 보행자는 시간이 지남에 따라 출현 여부가 달라지는 요인이므로, 이들을 구분해 처리하는 것이 장소 표현의 안정성을 높이는 데 기여할 수 있다.

그럼에도 불구하고 기존 연구는 시맨틱 클래스의 불확실성

Received : Sep. 5. 2025; Revised : Nov. 14. 2025; Accepted : Feb. 11. 2026
1. Assistant Professor, Mechanical Systems Engineering, Sookmyung Women's University, Seoul, Korea; Senior Research Engineer, Robotics LAB., Hyundai Motor Company, Uiwang, Korea (alexlee@sookmyung.ac.kr)
† Vice President, Corresponding author: Robotics LAB., Hyundai Motor Company, Uiwang, Korea (mecjin@gmail.com)



[Fig. 1] Activation distribution of the token [car] visualized with t-SNE on VLAD descriptors from the pitts30k dataset. Higher similarity corresponds to smaller cosine distance between CLIP image and text embeddings. The [car] token is concentrated in a few clusters, but the variance within clusters differs, indicating that activations are meaningful only in certain clusters

을 충분히 다루지 못했다. 보행자나 차량과 같은 동적 클래스를 단순히 제거하는 직관적 방식은, 특정 클래스의 발생이 장소 특성에 따른 조건부 분포를 따른다는 점을 간과한다. 예를 들어 car 클래스는 주차장에서는 빈번하지만 실내 환경에서는 거의 나타나지 않는다. [Fig. 1]에서 보듯 동일 토큰이라도 특정 클러스터에서는 유사도가 집중되어 안정적 단서로 작용하는 반면, 다른 클러스터에서는 분산되어 잡음처럼 보인다. 이는 동일 클래스의 활성화 분포가 장소 맥락에 따라 밀도와 분산이 달라짐을 의미하며, vision-language 모델을 통한 zero-shot 분석[8]에서도 관찰된다. 따라서 불확실성은 제거 대상이 아니라 장소 맥락에 따라 달라지는 단서로, 평균값 축약이 아닌 조건부 확률로 해석되어야 한다.

본 논문은 언어 기반 네비게이션의 맥락에서 이러한 불확실성을 재고한다. <주차장을 지나 엘리베이터 앞으로 이동하라>와 같은 지시는 특정 클래스의 존재와 공간적 위치를 전제로 하지만, 실제 장면에서 해당 클래스의 분포는 일정하지 않다. 따라서 본 논문은 유사도(similarity)나 신뢰도(confidence)와 같은 통계량에 의존해 불확실성을 단순화해 온 기존 접근에서 벗어나, 이를 장소-클래스에 대한 조건부 분포로 해석해야 한다는 문제의식을 제기한다.

2. 제안하는 접근과 실험

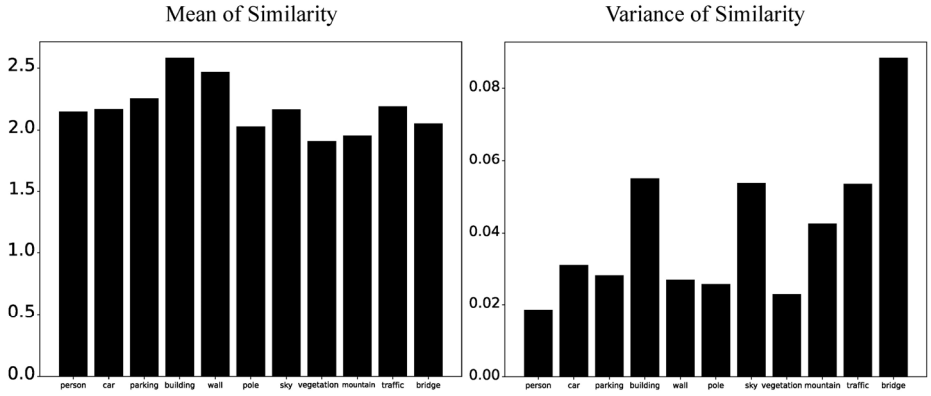
본 연구는 불확실성을 평균이나 분산과 같은 단일 통계량으로 축약하지 않고, 클래스-장소 조건부 분포로 해석하기 위한 분석 파이프라인을 설계하고 실증 관찰을 수행하는 데 목적이 있다. 새로운 알고리즘 제안이 아니라, 언어 기반 네비게이션에서 시맨틱 단서가 갖는 분포적 특성을 드러내고 이를 해석할 수 있는 절차를 제시하는 데 초점을 둔다.

분석 파이프라인은 세 단계로 요약된다. 첫째, Cityscapes[9]에서 정의된 시맨틱 클래스 토큰들(예: car, road, building)을 CLIP 텍스트 인코더로 임베딩 하고, 대응되는 이미지 임베딩과의 유사도를 계산한다. 둘째, 이미지 특징을 NetVLAD로 집약하여 전역 서술자(global descriptor)를 구성하고, 데이터베이스 전체에 대해 K-means로 장소 유형을 나타내는 클러스터를 생성한다. 본 연구에서는 장소의 분포적 다양성을 충분히 관찰하기 위해 64개 클러스터를 사용하였다. 셋째, 각 클러스터 내에서 클래스별 유사도 분포의 평균과 분산을 산출하고, t-SNE를 이용해 VLAD 설명자 분포를 시각화하여 클래스 활성화의 군집 특성과 비군집 특성을 판별한다. 이 절차를 통해 동일 클래스라도 클러스터에 따라 활성화 분포의 모양이 달라지는지를 관찰한다.

시각화 결과는 세 가지 관찰로 요약된다. 첫째, 일부 클래스가 소수의 클러스터에 강하게 집중되는 cluster-specific 패턴을 보임을 나타낸다. 교량(bridge), 신호기(traffic light) 등은 특정 장면군에서만 높은 유사도를 보이며, 해당 군 외부에서는 활성화가 희박해지는 경향이 관찰된다. 둘째, 건물(building)과 같이 많은 장면군에 널리 퍼진 non-cluster-specific 클래스 또한 존재한다. 이러한 클래스는 평균 유사도는 높을 수 있으나, 장소 판별력은 상대적으로 낮아지는 경향을 보인다. 셋째, 평균 유사도와 분산은 단순 비례하지 않는다. 특히 분산이 큰 클래스가 특정 장소 클러스터에서만 강하게 활성화되는 조건부 연관성을 지니는 사례가 반복적으로 확인된다. 이는 분산 자체가 정보량을 내포하는 지표가 될 수 있음을 시사하며, 분포의 모양을 고려하지 않은 유사도 기반 매칭이 구조적으로 한계를 가짐을 의미한다.

이러한 관찰은 단일 벡터 표현의 유사도 지표만으로 데이터베이스 전반의 top-k 관련성 보존을 보장할 수 없다는 이론적 지적과 합치한다[10]. 검색 표현을 하나의 벡터로 응축하는 순간 분포의 폭과 꼬리 행동이 담고 있던 장소 조건부 정보가 소거되기 쉽다. 또한 distributional reinforcement learning을 통해 환경 변동을 분포 자체로 모델링하는 접근[11]은 본 연구와 방향을 공유한다. 관찰된 고분산-집중형 패턴은 동적 클래스를 일괄 제거하는 전처리가 오히려 장소 구분에 유익한 맥락적 변동성을 삭제할 수 있음을 시사한다.

[Fig. 2]는 클래스-장소 조건부 분포의 비균일성을 정량적으로 보여준다. 좌측에서 다수 클래스의 평균 유사도 범위는 유사하지만, 우측에서 분산은 크게 갈라지므로 평균값만으로는 판별력이 부족한 상황이 발생한다. 저분산 클래스는 일부 클러스터에서 높은 유사도를 보이고 다른 클러스터에서는 약화되는 경향을 보여 장소 맥락에 결속된 단서로 해석되어야 한다. 반대로 고분산 클래스는 폭넓은 장소에서 안정적이지만 장소 구분력은 제한적일 수 있다. 따라서 동적 클래스를 일괄 제거하기보다, 클러스터 조건의 가중치 또는 prior로 분산 정보를 재활용하여야 한다.



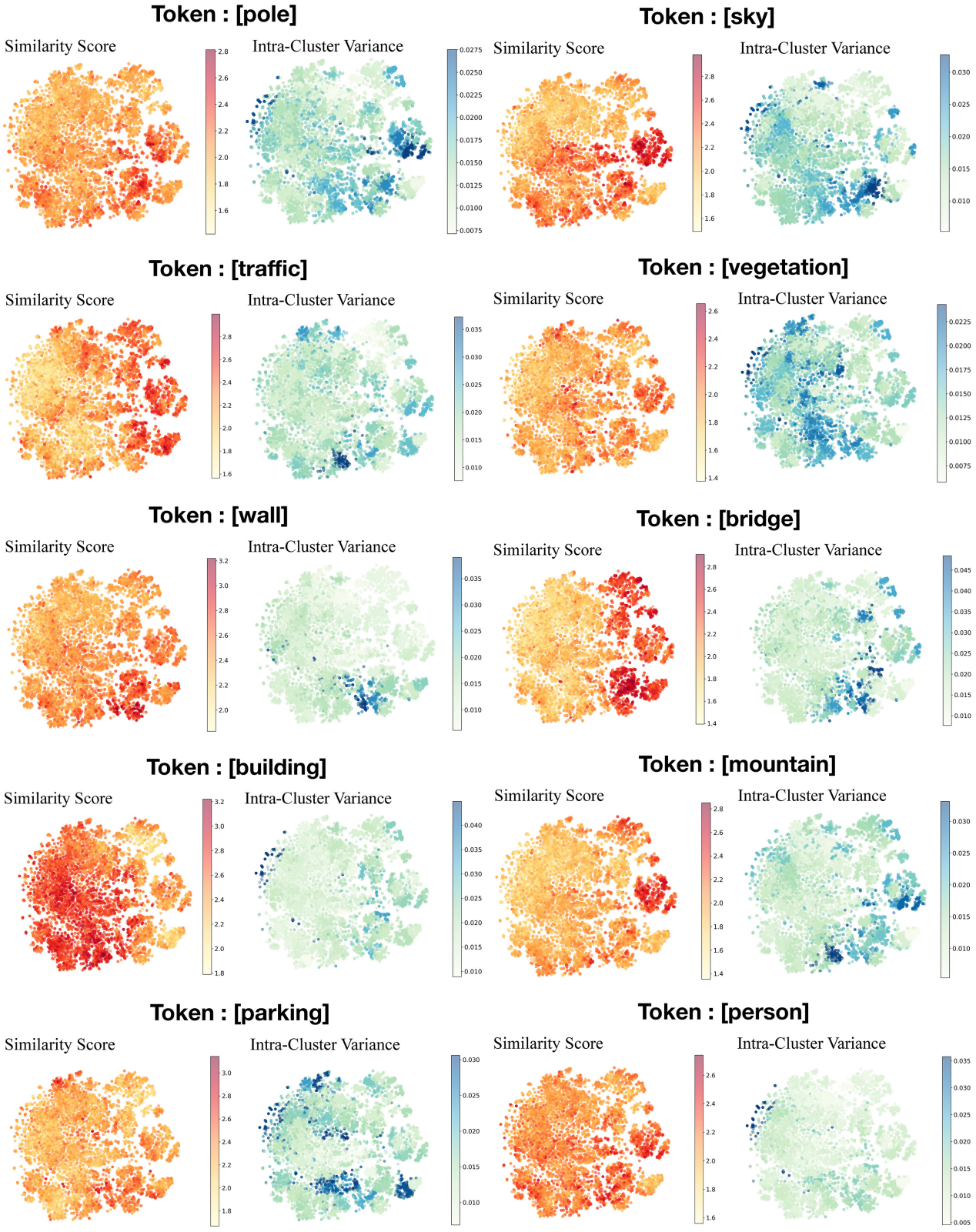
[Fig. 2] (left) Mean values and (right) variances of similarity from selected text classes on the pitts30k dataset. Higher variance indicates that a text class is more place-specific, when sufficient repeated measurements are obtained within the environment.

3. 논 의

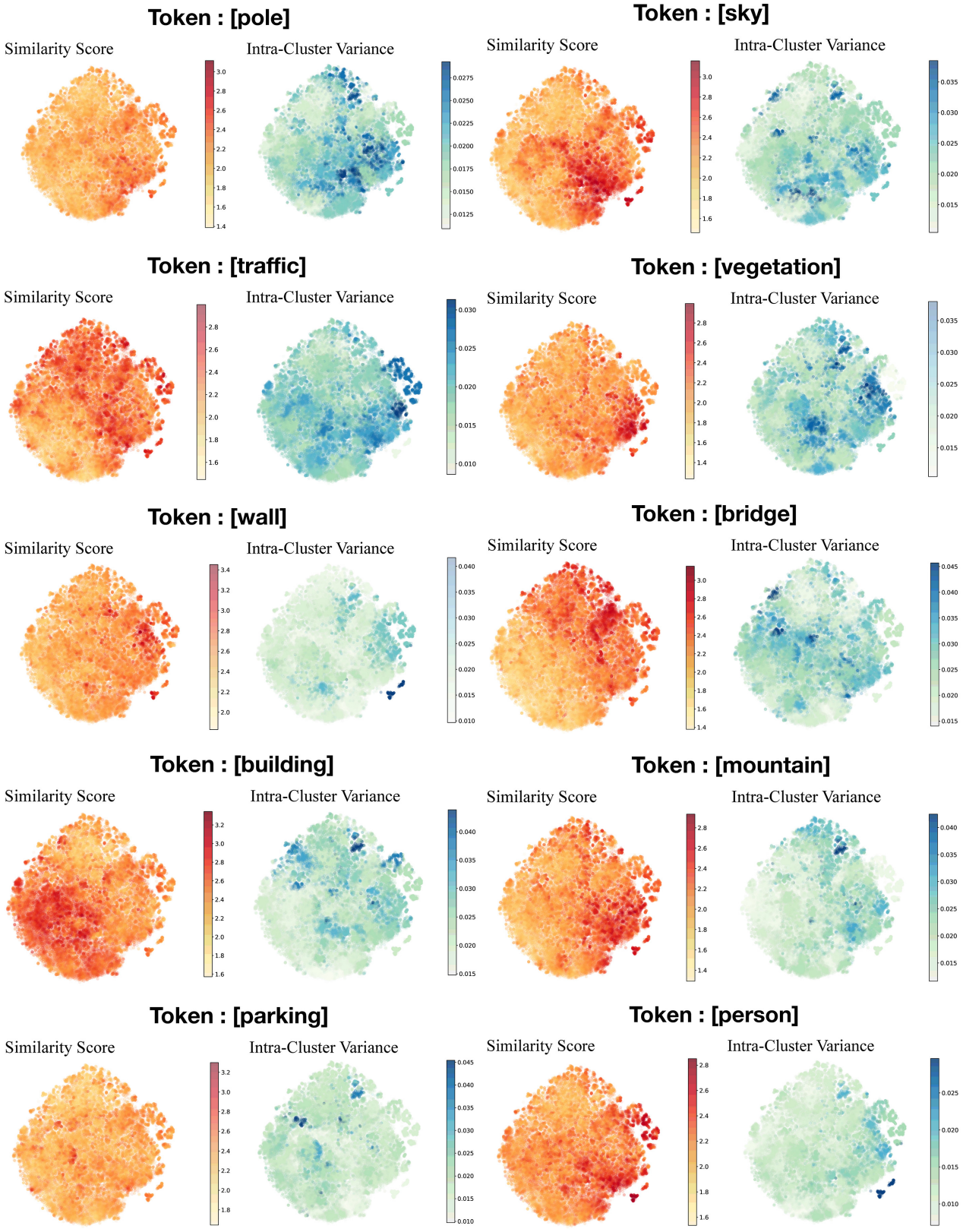
앞서 구성한 조건부 분포는 실제 네비게이션 및 장소 인식 과정에서 활용 가능한 사전 확률(prior)로 기능할 수 있다. 특히 후보 장소 k에 대해 관측된 클래스 활성화가 해당 장소의 조건부 분포와 어느 정도 일치하는지 평가하기 위해 앞서 제시한 최대우도를 사용할 수 있으며, 이는 기존 retrieval 점수와 결합하여 re-ranking을 수행하는 데 활용 가능하다. 이러한 절차는 입력 영상 혹은 점군에서의 시맨틱 측정이 특정 장소에서 통계적으로 타당한지 판단할 수 있게 하여, 언어 기반 네비게이션의 명령 수행 과정에서도 일관성 있는 판단 기준을 제공할 수 있다. 그러므로 표현 및 검색 단계에서 동적 클래스를 일괄 제거하기보다, 클러스터 조건의 가중치로 재활용하는 편이 합리적이다. 예컨대 전역 서술자를 활용할 때에 클러스터별 요약 통계(평균 혹은 분산 등)를 부가하거나, 멀티벡터/클러스터 조건 재랭킹 시에 사용하기 위해 분포 정보를 보존할 수 있다. 또한, 시맨틱 맵에는 클래스-장소 조건부 분포 요약을 저장하여 탐색 및 재방문 결정의 사전 정보(prior)로 활용하거나, 주행 시 위치 추정 안정화를 위한 정보로 활용하는 것이 유리하다. 구체적으로는, 클래스-장소 조건부 분포를 정규분포 형태로 근사한다면 이를 이용해 특정 장소 후보 k와 클래스 c에 대해 관측된 활성도 s_obs(c, k)에 대한 로그우도 $\log p(s_obs(c,k) \mid \mu(c,k), \sigma^2(c,k))$를 계산할 수 있다. 이 값은 관측 또는 연산된 시맨틱 활성도 정보가 해당 장소에서 통계적으로 타당한지를 정량적으로 나타내는 지표로 활용될 수 있다.

다만 본 연구에서는 다양한 로봇 임무요소에서 널리 통용되는 기저 모델이 없음을 감안하여, 위 모델링을 적용하였을 때의 전체 임무의 성능 향상보다는 클래스별 불확정성을 단순 적용하였을 때 점군 정합도의 변화를 제한적으로만 평가하였다. 시맨틱 클래스는 실험 단순화를 위해 [vegetation, ground, building] 세 가지로만 사용하였다. LiDAR를 사용하여 동일 장소 내 두 가지 다른 구역에서 각 여러 번 반복 측정을 수행하고, 첫 번째 구역에서 구성한 점군 지도의 시맨틱 클래스 활성도 s_obs를 ULIP2[12]을 통해 추정하였다. 이후 반복 측정 사이의 정합 오차를 기반으로 각 시맨틱 클래스에 대한 기댓값 및 표준편차를 측정하였으며, 두 번째 구역에서는 점군 지도를 작성 후 시맨틱 활성도 s_obs까지만을 추정하여 이를 기반으로, 첫 번째 구역에서 산출된 s_obs와 정합 오차와의 상관관계 중 그 평균값만을 사용하여 각 3차원 측정점에 중요도 가중치를 부과하였다. 그 결과, RMSE가 0.1490에서 0.1429로 감소하였으며, 현 실험과 같은 일대일 정합이 아닌 일대 다 정합의 경우에는 위에 제시된 것과 같은 최대우도 방법으로 표준편차까지 활용할 수 있을 것이다. 비록 이 실험이 단순한 가중치 방식이며 분산 통계량을 직접 활용한 것은 아니나, 단순한 matching score 평가 만으로 정합 점수 개선이 이루어졌으므로, 추후 분포 정보를 활용할 경우 실제 정합 품질이 개선될 여지가 있음을 알 수 있다.

본 연구는 분포 기반 해석의 가능성을 확인하는 초기 단계 연구로, 활용한 이미지 기반 zero-shot segmentation 모델의 일반화 능력이 완벽하지 않아 다양한 환경과 데이터셋에 대한 확장 검증이 필요하며, 최대우도 기반 평가 지표를 활용하여 분포 정보를 기반으로 re-ranking을 수행하는 등 실제 네비게이션과 관련된 장소인식 등의 지표 향상에 연결하는 추후 연구가 요구된다. 그러므로 향후 연구에서는 본 논문에서 제안한 조건부 분포 기반 최대우도 지표를([Fig. 3], [Fig. 4]) 점군 / 이미지 기반 장소 인식 파이프라인에 통합하여, 다양한 네비게이션 벤치마크에서의 성능 개선 여부를 체계적으로 평가하여야 한다.



[Fig. 3] Activation distribution of selected classes text from CityScapes[9] class definitions, visualized using t-SNE on VLAD descriptors from the pitts40k dataset's training sequence



[Fig. 4] Activation distribution of selected classes text from CityScapes[9] class definitions, visualized using t-SNE on VLAD descriptors from the pitts250k dataset's training sequence

References

[1] O. Siméoni et al., "DINOv3," *arXiv preprint arXiv*:2508.10104, Aug., 2025, DOI: 10.48550/arXiv.2508.10104.

[2] M. Caron et al., "Emerging properties in self-supervised vision transformers," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Montreal, QC, Canada, pp. 9650-9660, 2021, DOI: 10.1109/ICCV48922.2021.00951.

[3] A. Radford et al., "Learning transferable visual models from natural language supervision," in *Proc. Int. Conf. Mach. Learn. (ICML)*, Virtual, vol. 139, pp. 8748-8763, 2021, DOI: 10.48550/arXiv.2103.00020.

[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*, vol. 9, no. 2, pp. 1286-1293, Feb., 2024, DOI: 10.1109/LRA.2023.3343602.

[5] H. Jégou, M. Douze, C. Schmid, and P. Pérez, "Aggregating local descriptors into a compact image representation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, San Francisco, CA, USA, pp. 3304-3311, 2010, DOI: 10.1109/CVPR.2010.5540039.

[6] R. Arandjelović, P. Gronat, A. Torii, T. Pajdla, and J. Sivic, "NetVLAD: CNN architecture for weakly supervised place recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Las Vegas, NV, USA, pp. 5297-5307, 2016, DOI: 10.1109/CVPR.2016.572.

[7] V. Paolicelli, A. Tavera, C. Masone, G. Berton, and B. Caputo, "Learning semantics for visual place recognition through multi-scale attention," in *Proc. Int. Conf. Image Analysis and Processing (ICIAP)*, Lecce, Italy, LNCS 13232, Springer, pp. 454-466, May, 2022, DOI: 10.1007/978-3-031-06430-2_38.

[8] J.-H. Choi, H.-J. Baek, C.-S. Park, and I. Kim, "Utilizing AI Foundation Models for Language-Driven Zero-Shot Object Navigation Tasks," *J. Korea Robot. Soc. (KRS Journal)*, vol. 19, no. 3, pp. 293-310, Sep., 2024, DOI: 10.7746/jkros.2024.19.3.293.

[9] M. Cordts et al., "The Cityscapes dataset for semantic urban scene understanding," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Las Vegas, NV, USA, pp. 3213-3223, 2016, DOI: 10.1109/CVPR.2016.350.

[10] O. Weller, M. Boratko, I. Naim, and J. Lee, "On the Theoretical Limitations of Embedding-Based Retrieval," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2026, DOI: 10.48550/arXiv.2508.21038.

[11] V. M. Tran and G.-W. Kim, "Mapless Navigation with Distributional Reinforcement Learning," *J. Korea Robot. Soc. (KRS Journal)*, vol. 19, no. 1, pp. 92-97, Mar., 2024, DOI: 10.7746/jkros.2024.19.1.092.

[12] L. Xue et al., "ULIP-2: Towards Scalable Multimodal Pre-training for 3D Understanding," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Seattle, WA, USA, pp. 27091-27101, 2024, DOI: 10.1109/CVPR52733.2024.02558.

이 준 호

2017 B.S, Mechanical Engineering, KAIST

2023 Ph.D., Robotics Program, Civil and Environmental Engineering, KAIST

2023-2025 Robotics Lab, Hyundai Motor Company

2025~current Mechanical Systems Engineering, Sookmyung Women's University

관심분야: Spatial Perception, Robotics

현 동 진

2006 B.S, Mechanical Engineering, Seoul National University

2007 M.S. Computational Mechanics, University of Michigan

2012 Ph.D., Control/Dynamics, UC Berkeley

2012~2014 Postdoctoral Associate, Massachusetts Institute of Technology

2014~2026 Robotics LAB., Hyundai Motor Company

관심분야: Legged Robotics, Wearable Robotics, Motion Control