Journal of Institute of Control, Robotics and Systems (2025) 31(8):882-887
https://doi.org/10.5302/J.ICROS.2025.25.0161 ISSN:1976-5622 eISSN:2233-4335

가림 상황에 강인한 랜덤 마스킹을 이용한 대조학습 기반 교차시점 이미지 리트리벌 프레임워크

Occlusion-robust Cross-view Geo-image Retrieval via Contrastive Learning and Random Masking

이 승 희*
(Seunghee Lee[1,*])
[1]Agency for Defense Development (ADD)

Abstract: Cross-view geo-image retrieval (CVIR) aims to retrieve geographically corresponding satellite images given ground-view images, playing a vital role in various applications such as autonomous driving and unmanned aerial vehicles (UAVs). Although prior studies have mainly focused on mitigating the viewpoint gap through geometric transformation techniques, they have largely overlooked real-world challenges such as occlusions caused by dynamic objects and infrastructure. To address this issue, we propose a contrastive learning framework that introduces random masking to effectively simulate diverse occlusion scenarios during training and enhance the model's robustness. By applying random masking to the input images during training, the proposed method reduces the model's reliance on local visual features and enables it to learn more robust features. Our approach requires no transformation or alignment processes, providing a straightforward and effective solution for the occlusion gap. Extensive experiments on standard benchmarks demonstrate that the proposed framework achieves state-of-the-art performance while maintaining robustness under various occlusion conditions.

Keywords: cross-view image retrieval, random masking, contrastive learning, occlusion robustness

I. 서론

정밀한 위치 추정 기술은 자율주행, 무인항공기(UAV) 운용, 군사 감시, 로봇 네비게이션 등 다양한 응용 분야에서 핵심적인 역할을 한다[1,2]. 기존의 대부분의 위치 기반 서비스는 GNSS (Global Navigation Satellite System)에 의존하고 있지만, 도심 지역의 고층 건물이나 지하, 실내 공간, 혹은 군사 작전 지역과 같은 GNSS-denied 환경에서는 신호 차단 및 왜곡으로 인해 정확도가 급격히 저하되는 문제가 있다. 이러한 한계를 극복하기 위해 최근에는 위성 이미지와 지상 이미지 간의 시각적 대응 관계를 기반으로 위치를 추정하는 교차시점 지오이미지 로컬라이제이션(cross-view geo-image localization, CVGL) [3-5] 기술이 주목받고 있다. CVGL의 핵심 과제는 교차시점 지오이미지 리트리벌(cross-view geo-image retrieval, CVIR) 기술로, 입력으로 주어진 지상 이미지(예: 거리에서 촬영된 사진)에 대해 동일 위치에서 촬영된 위성 이미지를 검색하는 문제이다. CVIR의 두 입력 이미지 도메인이 서로 완전히 다른 시점(viewpoint)에서 촬영되었기 때문에 시각적 형태, 구조, 배경 정보가 크게 달라지는 특성이 있다. 이러한 시점 차이(viewpoint gap)는 이미지 간 표현 불일치의 주요 원인으로, CVIR의 성능을 제한하는 핵심 요인 중 하나로 지적되어 왔다.

이를 해결하기 위해 기존 연구[6-15]들은 다양한 기하학적 정렬 기법이나 attention 기반[16] 구조를 활용하여 위성 이미지와 지상 이미지 간의 표현 간극을 줄이고자 하였다. 대표적으로 polar transformation [14]이나 spherical projection [15]을 통해 서로 다른 시점에서 촬영된 이미지를 정렬한 후, deep feature matching을 수행하는 구조가 널리 활용되었다. 또한 최근에는 복잡한 정합 과정 없이 도메인 간 표현을 효과적으로 정렬할 수 있는 대조학습(contrastive learning) 기반의 프레임워크도 제안되었다.

그러나 대부분의 기존 연구들[7,9-13]은 viewpoint gap의 완화에 집중하고 있으며, 실제 환경에서 빈번하게 발생하는 가림(occlusion) 문제는 거의 고려하지 않았다. 실제 거리 이미지에서는 차량, 나무, 가로등, 행인 등 다양한 동적 또는 정적 객체에 의해 장면의 일부분이 가려질 수 있으며, 이는 동일 위치에서 촬영된 이미지 쌍이라 하더라도 시각적으로 상이한 특징을 가지게 만든다. 이러한 occlusion gap은 모델이 특정 객체나 지역에 과도하게 의존할 경우 일반화 성능 저하로 이어질 수 있으며, 실제 응용에서는 매우 치명적인 약점이 된다.

본 연구에서는 이러한 문제를 해결하기 위해, 랜덤 마스킹(random masking) 기반의 학습 전략을 통합한 새로운 대조학습 프레임워크를 제안한다. 학습 과정에서 입력 이미지의 일부를 마스킹함으로써 다양한 가림 상황을 경험하도록 유도하며, 이를 통해 특정 로컬 피처에 영역에 과적합되지 않고 전체적인 구조적 맥락을 이해하는 표현을 학습할 수 있도록

* Corresponding Author
Manuscript received May 26, 2025; revised June 10, 2025; accepted July 5, 2025
이승희: 국방과학연구소 연구원(seungheelee@add.re.kr, ORCID 0009-0007-5274-3040)
※ 이 논문은 2025년 정부(방위사업청)의 재원으로 수행된 연구임.

한다. 또한 마스킹 비율을 선형적으로 증가시키는 커리큘럼 방식의 학습을 도입하여, 초기에는 전체 이미지 정보를 바탕으로 안정적으로 학습하고, 이후에는 점진적으로 제한된 정보에서도 강건한 표현을 유지할 수 있도록 구성하였다. 특히 본 연구에서는 같은 도메인 내에서 마스킹된 이미지와 원본 이미지를 비교하는 self-modality learning 과 서로 다른 도메인 간에 원본 또는 마스킹 이미지를 교차 비교하는 cross-modality learning 학습을 활용하여, 안정적이고 일관된 표현을 학습하며, 시점 차이와 가림 문제를 동시에 극복할 수 있다. 제안 방법은 별도의 복잡한 변환이나 정합 모듈 없이, 단일 encoder 기반의 간결한 구조로 구현 가능하며, VIGOR [17] 및 CVUSA [5] 와 같은 CVGL 표준 벤치마크 데이터셋에서 기존 연구 대비 우수한 성능과 강건성을 입증하였다.

II. 관련 연구

초기의 기존 연구들[6-15]은 다양한 시각적 정렬 기법을 제안해왔으며, 특히 polar transformation, spherical projection과 같은 기하학적 변환 방식과 attention 기반의 특징 집계 구조가 널리 활용되었다. 대표적으로 [14]는 위성 이미지를 polar transform을 통해 ground-view 시점과 유사하게 재투영하고, 이에 SAFA (Spatial-Aware Feature Aggregation) 모듈을 결합하여 도메인 간 표현 차이를 완화하였다. 이후 TransGeo [16]는 Vision Transformer 기반 backbone에 멀티스케일 attention 및 zoom 모듈을 결합하여 viewpoint 변화에 더욱 유연하게 대응하는 구조를 제시하였으며, SAIG-D [18]는 MLP-Mixer 기반의 백본을 활용하여 구조적 단순성과 표현 성능을 동시에 확보하였다. GeoDTR [12]은 geometric layout 정보를 disentangle하는 구조로 표현력 향상을 도모하였다. 이러한 방법들은 시점 차이 감소에는 효과적이었으나, 대부분 복잡한 네트워크 구조나 추가적인 처리 모듈이 요구된다는 단점이 있다.

최근에는 복잡한 변환이나 정합 구조 없이도 효율적인 표현 학습을 위해 대조학습기반[19-21] 접근이 주목받고 있다. 대조학습은 복잡한 정렬 모듈 없이도 표현 공간에서 이미지 간 유사도를 직접적으로 학습할 수 있다. Sample4Geo [20]는 InfoNCE 손실을 기반으로 하여 양성 쌍 간의 거리를 줄이고 음성 쌍 간 거리를 벌리도록 학습하였으며, 초기에는 GPS 기반의 근접 샘플링, 이후에는 cosine 유사도를 기반으로 한 hard negative mining 전략을 통해 효과적인 표현 학습을 구현하였다. 해당 프레임워크는 기존 기법과 유사하거나 더 나은 성능을 달성하였다. 하지만 Sample4Geo 역시 viewpoint gap의 완화에 주로 집중되어 있고, occlusion 상황에 대한 대응은 제한적이다.

컴퓨터 비전 분야에서는 이러한 문제를 완화하기 위해 무작위 마스킹(Random Masking)을 적용한 self-supervised 학습 전략이 다양한 형태로 연구되어 왔다. 대표적으로 MAE [22]는 입력 이미지의 일정 비율을 마스킹한 후 복원하는 과정을 통해 강건한 표현을 학습하였으며, SimMIM [23], BEiT [24] 등도 유사한 구조를 통해 특정 지역에 대한 과적합을 방지하고 전역적인 scene 이해를 유도하였다. 그러나 이러한 마스킹 기반 전략은 대부분 동일시점(viewpoint) 학습에 한정되어 있으며, cross-view 환경에서 occlusion robustness를 확보하기 위한 구조적인 응용은 시도된 바 없다. 본 연구는 기하 변환 없이 마스킹 기반 증강과 대조학습만으로도 occlusion에 강건한 표현을 학습할 수 있는 새로운 프레임워크를 제안한다. Self-modality 및 cross-modality 손실을 결합하여, 다양한 가림 조건에 적응 가능한 표현 학습이 가능하다.

III. 제안 방법

1. 전체 프레임워크 구조

전체 프레임워크 구조는 그림 1과 같이, 위성 이미지(I_S)와 지상 이미지(I_G)를 입력으로 하여 각각 무작위 마스킹을 적용한 변형 이미지(I_{SM}, I_{GM})를 추가 생성한다. 원본 및 마스킹 이미지들은 공유 인코더(shared encoder) $M(\cdot)$를 통해 임베딩 벡터(F_S, F_G, F_{SM}, F_{GM})로 변환된다. 이 임베딩들은 대조학습 기반의 손실 함수를 통해 학습되며, 동일한 위치에서 촬영된 위성-지상 이미지 쌍이 가림 여부에 관계없이 유사한 표현 공간으로 정렬되도록 유도한다.

2. 랜덤 마스킹 기법

랜덤 마스킹은 입력 이미지의 일부 영역을 확률적으로 제거하여 모델이 특정 영역에 과도하게 의존하지 않고 전체적인 맥락 정보를 학습할 수 있도록 한다. 각 입력 이미지 $I \in \mathbb{R}^{H\times W\times C}$ 는 높이 H, 너비 W, 채널 수 C를 가진 RGB 이미지로, 이를 일정 크기의 patch 로 분할한 후, 무작위로 선택된 일정 비율의 patch에 마스크를 적용한다. 각 patch는 확률 p에 따라 제거되며, 최종 마스킹된 이미지 I^m는 그림 2와 같으며 다음과 같이 계산된다.

$$I^m = I \odot \big(1 - M(p)\big), where\ M(p) \in \{0,1\}^{H\times W} \quad (1)$$

여기서 마스크 행렬 $M(p)$는 각 픽셀이 마스킹될 확률 p에 따라 무작위로 생성되는 이진 행렬이다. 초기 학습 단계에서는 마스킹 비율 p를 작게 설정하여 안정적인 학습을 유도하고, 학습이 진행됨에 따라 점진적으로 p를 증가시키는 커리큘럼 학습 전략을 도입하여 다양한 가림 조건에 점진적으로 적응할 수 있도록 유도하였다. 이러한 마스킹은 매 epoch마다 새로운 패턴으로 적용되어 정적인 영역 또는 특정 구조에 대한 과적합을 방지하고, 다양한 가림 조건에서 견고한 표현을 학습하는 데 기여한다.

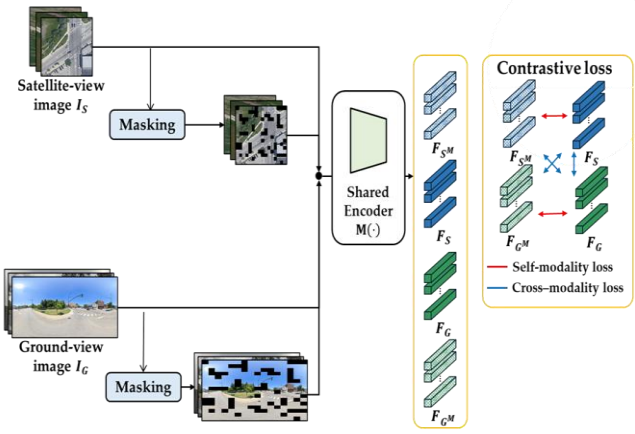

그림 1. 전체 네트워크 구조.
Fig. 1. The overall framework.

그림 2. 랜덤 마스킹 적용한 이미지 예시.
Fig. 2. Examples of images with random masking applied.

3. 대조학습 목표 및 손실 함수

제안하는 프레임워크는 InfoNCE 손실 함수를 기반으로 대조학습을 수행한다. 대조학습은 positive 쌍의 표현 유사도는 높이고 negative 쌍의 표현 유사도는 낮추는 방향으로 표현 공간에서 특징을 학습하는 self-supervised 학습 기법이다. 이는 양성 쌍(positive pair)의 유사도를 극대화하고 음성 쌍(negative pairs)과의 유사도를 최소화하는 방식으로 정의된다. InfoNCE 손실은 식 (2)와 같다:

$$\mathcal{L}(f_q, f_r) = -\frac{1}{B}\sum_{i=1}^{B} \log \frac{\exp\left(f_{q_i} \cdot \frac{f_{r_i}}{\tau}\right)}{\sum_{j=1}^{B} \exp\left(f_{q_j} \cdot \frac{f_{r_j}}{\tau}\right)} \quad (2)$$

여기서 f_q과 f_r 은 각각 쿼리와 레퍼런스 이미지의 임베딩이며 τ 는 온도 파라미터, B는 쿼리와 레퍼런스의 배치수이다. 전체 손실함수는 식 (3)과 같이 정의되며, 여기서 w_1와 w_2는 각각 self-modality 및 cross-modality 손실 항목의 상대적인 기여도를 조정하는 하이퍼파라미터이다.

$$\mathcal{L}_{total} = \mathcal{L}_{base} + w_1 \cdot \mathcal{L}_{self-modality} + w_2 \cdot \mathcal{L}_{cross-modality} \quad (3)$$

3.1. $\mathcal{L}_{base} = \mathcal{L}(f_g, f_r)$

여기서, $\mathcal{L}_{base}$는 서로 다른 두 도메인의 원본 이미지들을 비교함으로써 가장 기본적인 positive pair 간의 특징을 임베딩 스페이스에서 가까이 위치시킬 수 있도록 학습하게 해준다.

3.2. $\mathcal{L}_{self-modality} = \mathcal{L}(f_g, f_{gm}) + \mathcal{L}(f_s, f_{sm})$

Self-modality 손실함수는 원본 이미지와 마스킹 이미지 간의 유사성을 학습하며, 특정 로컬 영역에 집중하지 않고 다양한 피처를 학습함으로써 동일 도메인 내 일관된 특징을 학습하여 부분적인 가림에도 강인한 특징을 뽑아낼 수 있게 한다.

3.3. $\mathcal{L}_{cross-modality} = \mathcal{L}(f_g, f_{sm}) + \mathcal{L}(f_s, f_{gm})$

Cross-modality 손실함수는 서로 다른 도메인 간의 원본과 마스킹 쌍 간의 유사성을 학습하며, 가림이 존재하는 서로 다른 도메인의 쌍들도 서로 가깝게 학습되도록 유도함으로써 일부 로컬 피처에 오버피팅되지 않고 다양한 지형적 의미론적 특징을 학습할 수 있게 한다.

제안한 self-modality loss는 같은 도메인 내에서 원본 이미지와 마스킹된 이미지 간의 표현 일관성을 확보하여 이미지의 일부분이 가려진 경우에도 전체적인 구조적 맥락을 유지하는 강건한 특징 표현을 학습하도록 유도한다. 반면 cross-modality loss는 서로 다른 도메인 간 원본 및 마스킹 이미지를 교차적으로 비교하여 도메인이 다르거나 viewpoint 차이가 존재하더라도 이미지 간 표현을 일관되게 유지할 수 있도록 학습한다. 따라서 두 손실 함수의 조합은 모델이 다양한 형태와 수준의 occlusion 상황에서 특정 로컬 영역이나 객체에 민감하게 의존하지 않고, 보다 일반적이고 견고한 표현을 학습할 수 있도록 한다. 전체 손실함수 식 (3)은 원본과 마스킹 이미지 간의 일관성 있는 표현을 학습하여 occlusion 이 존재하는 환경에서도 견고하고 일반화된 리트리벌 성능을 달성하는 데 기여한다.

IV. 실험 및 결과

1. 실험셋팅

1.1 데이터셋

모델의 성능을 평가하기 위해 두 가지 대표적인 CVIR 벤치마크 데이터셋인 VIGOR [17]와 CVUSA [5]를 사용하였다. VIGOR 데이터셋은 뉴욕, 시애틀, 샌프란시스코, 시카고 등 4개 도시에서 수집된 90,618장의 위성 이미지와 105,214장의 지상 이미지로 구성된 대규모 데이터셋으로 정밀 정합을 어렵게 만드는 semi-positive 샘플이 존재한다. CVUSA 데이터셋은 미국 전역에서 수집된 35,532 쌍의 위성 및 지상 이미지로 구성된 학습 데이터와 8,884 쌍의 테스트 데이터로 구성되어 있으며, 위성 이미지는 750×750, 지상 이미지는 224×1232 해상도로 북향 정렬(North-aligned)된 구조를 가지며 one-to-one 매핑 구조로 되어 있다.

1.2 Implementation details

제안된 모델은 PyTorch로 구현되었으며, 모든 실험은 NVIDIA H100 GPU를 사용하여 수행하였다. ConvNeXt-Base [25]를 encoder 구조로 사용하였으며, 위성 이미지와 지상 이미지 모두 동일한 encoder를 weight-sharing 하였으며, 이미지의 입력 크기는 모두 384×384로 통일하였고, 배치 크기는 128, 학습 epoch 은 40으로 설정하였다. 최적화 방법으로 AdamW optimizer를 사용하였고, 학습 초기 learning rate는 0.001 로 설정하였다. 또한 cosine decay scheduler를 적용하였으며, warmup epoch 은 1로 지정하였다. 마스킹 비율은 학습 초기 0에서 시작하여 최종적으로 0.9까지 점진적으로 증가하도록 설정하였고, 입력 이미지는 8×8크기의 patch로 분할하여 마스킹을 수행하였다. 본 모델의 랜덤 마스킹과 self-/cross-modality 손실 기법은 오직 학습 단계에서만 적용되며, inference 시에는 기존 Sample4Geo 모델과 동일한 구조 및 연산 방식을 사용하여 추가 연산 비용(FLOPs) 등의 변화는 없다.

1.3 평가 지표

Recall@K(R@K)는 쿼리 이미지에 대해 모델이 반환한 K개의 결과 중에 정답이 포함되어 있으면 성공으로 간주하며, 전체 쿼리 중 정답이 포함된 비율로 계산한다. 본 연구에서는 R@1, R@5, R@10과 각 쿼리 이미지에 대해 전체 위성 이미지 데이터베이스를 대상으로 검색했을 때, 전체 위성 이미지 수의 상위 1% 이내에 정답 이미지가 포함되는 비율인 R@1%를 함께 사용하였다. VIGOR 데이터셋에는 각 쿼리 이미지에 대해 하나의 정답 외에도 여러 개의 semi-positive 이미지가 함께 존재하기에, 정답 또는 semi-positive 중 하나라도 검색 결과 상위 1개에 포함될 경우 정답으로 간주하는 Hit rate 지표를 함께 사용하였다.

2. 정량적 성능 비교

표 1은 VIGOR 데이터셋에서 표 2는 CVUSA 데이터셋에서

표 1. VIGOR 데이터셋에서의 정량적 실험 결과.
Table 1. Quantitative comparison on VIGOR.

Approach	VIGOR				
	R@1	R@5	R@10	R@1%	H.R
SAFA [14]	33.93	58.42	68.12	98.24	36.87
TransGeo [16]	61.48	87.54	91.88	99.56	73.09
Sample4Geo [20]	77.63	95.54	97.12	99.64	89.59
Ours	**79.84**	**96.44**	**97.74**	**99.70**	**91.32**

표 2. CVUSA 데이터셋에서 정량적 실험 결과.
Table 2. Quantitative comparison on CVUSA.

Approach	CVUSA			
	R@1	R@5	R@10	R@1%
SAFA [14]	89.84	96.93	98.14	99.64
TransGeo [16]	94.08	98.36	99.04	99.77
Sample4Geo [20]	98.68	99.68	**99.78**	99.87
Ours	**98.75**	**99.69**	99.76	**99.89**

제안한 방법과 기존 주요 기법들과의 성능 비교 결과를 보여준다. 모든 데이터셋에서 제안하는 모델은 기존 sample4 geo를 포함한 기존 연구들에 비해 우수한 성능을 보였다.

3. 정성적 성능 비교

정량적 지표 외에도, 제안된 기법의 표현 학습 특성과 시각적 강건성을 보다 직관적으로 확인하기 위해 정성적 분석을 수행하였다. 구체적으로는 쿼리 지상 이미지에 대해 상위 5개 검색 결과 (top-5 retrieval results)를 시각화하고, Sample4 Geo [20] 와의 결과를 그림 3과 같이 비교하였다. 그 결과, 자동차나 나무 등의 배경 변화가 있을 수 있는 쿼리에 대해

그림 3. Top-5 retrieval 결과.
Fig. 3. Examples of top-5 retrieval results.

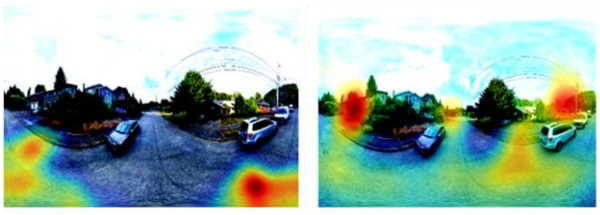

그림 4. Sample4Geo (왼쪽)와 제안방법(오른쪽)의 주의 집중 영역 시각화 예시. 기존 모델은 일부 local feature에 집중하지만, 제안한 모델은 다양한 local feature를 활용한다.
Fig. 4. Visualization of attention maps for Sample4Geo(left) and our method(right). Ours shows a more spatially distributed activation pattern than Sample4Geo.

오답을 반환하는 경우가 더 많았던 반면, 제안 기법은 정확한 위성 이미지를 반환하는 경향을 보였다. 이는 랜덤 마스킹을 통해 로컬 정보에 의존하지 않고 전체 맥락을 활용한 표현을 학습하였기 때문으로 해석된다.

또한, 그림 4는 attention map을 통해 GradCAM 결과를 보여준다. 모델이 집중하는 곳은 빨간색으로 덜 집중하는 곳은 파란색으로 표현된다. 제안하는 모델이 Sample4Geo에 비해 더 넓은 범위의 시각적 단서에 주목하며 특징을 추출함을 확인할 수 있다

4. 추가 실험

4.1 각 항목 별 성능 비교

제안한 프레임워크를 구성하는 각 손실 항목의 기여도를 정량적으로 분석하기 위해, 체 손실 함수에서 개별 항목을 제거하거나 조합을 달리하여 학습한 후, VIGOR 데이터셋에서 비교하였다. 표 3에서와 같이, cross modality loss를 추가하면 서로 다른 도메인 간 일관성 있는 표현 학습이 가능해져 전체적인 recall 성능이 향상되었고, self-modality loss는 같은 도메인 내에서 마스킹된 이미지와의 일관성을 강화하여, 전체 성능의 안정성을 높이는 데 기여하였다. 모든 손실 항목을 통합 적용한 경우 가장 높은 성능을 기록하였으며, 이는 각각의 손실 항이 상호보완적으로 작용함을 보여준다.

표 3. VIGOR 데이터셋에서의 각 항목 별 정량적 실험 결과. 여기서 V는 vanilla, C는 cross-modality, S는 self-modality, H.R 은 hit rate를 의미한다.
Table 3. Quantitative comparison on VIGOR. In the approach, V means 'vanilla', C means 'cross-modality', S means 'self-modality', and H.R. means 'hit rate'.

Approach			VIGOR				
V	C	S	R@1	R@5	R@10	R@1%	H.R
O			77.63	95.54	97.12	99.64	89.59
O	O		78.09	96.01	97.43	99.69	90.24
O		O	79.60	96.23	97.61	99.69	90.76
O	O	O	79.84	96.44	97.74	99.70	91.32

4.2. 가림상황에서의 성능 비교

가림 상황에서 모델의 강건성을 평가하기 위해, VIGOR 데이터셋의 테스트 이미지 전체에 instance segmentation 기반의 data augmentation 기법[26]을 활용하여 차량과 사람 등 실제 도심 환경에서 자주 등장하는 dynamic object를 지상 입력 이미지에 삽입한 추가적인 가림 상황을 구현하였다. 각 이미지당 추가된 dynamic object의 개수는 최소 2개에서 최대 10개까지 다양하게 설정되었으며, 객체의 크기와 위치 또한 무작위로 선정하여 이미지 내의 다양한 영역에서 현실적인 가림 상황이 발생하도록 구성하였다. 그림 5는 dynamic object 5개가 삽입된 테스트 이미지 예시이며, 이렇게 생성된 이미지들은 학습에는 사용되지 않고 오직 occlusion 정도에 따른 성능 평가만을 위한 inference 용으로 활용되었다.

그림 6과 7은 각각 dynamic object의 개수가 증가함에 따라 모델의 리트리벌 성능이 어떻게 변화하는지 R@1 및 R@5 기준으로 나타낸 것이다. 제안한 방법과 비교모델(Sample4Geo) 모두 가림 객체 수가 증가할수록 리트리벌 성능이 점차 저하되었으나, 제안한 모델은 Sample4Geo 대비 성능 저하가 상대적으로 완만하였으며, 다양한 수준의 가림 상황에서도 일관되고 강건한 리트리벌 성능을 유지함을 확인할 수 있었다. 이는 본 논문에서 제안한 랜덤 마스킹 기반의 self-modality 및 cross-modality 손실 함수를 이용한 기법이 실제적인 occlusion 환경에서 강건한 특징 표현을 효과적으로 학습하는 데 기여하였음을 명확히 보여준다.

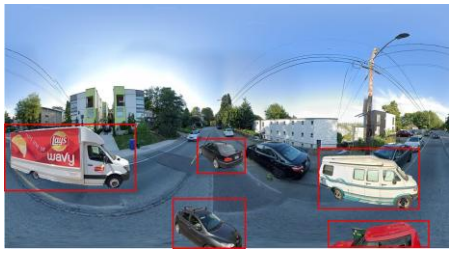

그림 5. Dynamic object 가 삽입된 테스트 이미지 예시.
Fig. 5. An example of an occlusion scenario image.

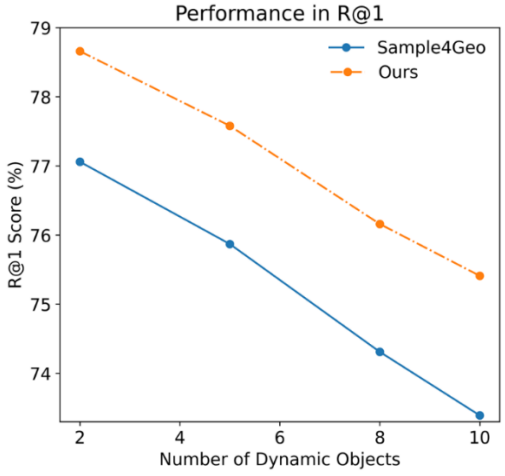

그림 6. Dynamic object 수 증가에 따른 R@1 성능 변화 비교.
Fig. 6. Comparison of R@1 performance under increasing numbers of dynamic objects.

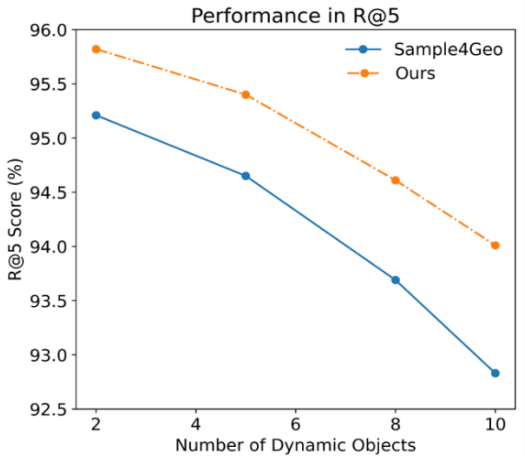

그림 7. Dynamic object 수 증가에 따른 R@5 성능 변화 비교. 제안한 방법은 Sample4Geo 대비 occlusion 상황에서 더 높은 성능을 유지함.
Fig. 7. Comparison of R@5 performance under increasing numbers of dynamic objects. Ours shows more robust performance against occlusion than Sample4Geo.

V. 결론

본 논문에서는 기하 변환 없이도 occlusion gap을 완화할 수 있는 랜덤 마스킹 기반 대조학습 프레임워크를 제안하였다. 제안한 방법은 입력 이미지의 일부분을 무작위로 제거하고, self-modality 및 cross-modality 간의 대조학습 손실을 함께 활용하여 특정 로컬 피처에 의존하지 않는 강건한 표현을 학습하도록 설계되었다. 또한 마스킹 비율을 학습 epoch에 따라 점진적으로 증가시키는 커리큘럼 전략을 통해 다양한 가림 조건에 적응할 수 있도록 하였다. VIGOR와 CVUSA 등 주요 벤치마크 데이터셋에서의 실험에서, 제안한 방법은 기존 최신 기법 대비 우수한 성능을 보였으며, 특히 occlusion 시뮬레이션 환경에서도 우수한 일반화 성능을 입증하였다. 본 연구는 복잡한 기하 변환 없이 간결한 구조만으로도 occlusion gap 문제를 효과적으로 완화할 수 있음을 실증하였으며, 향후 다양한 occlusion-aware 학습 전략이나 입력 모달리티 확장에 응용 가능할 것으로 기대된다. 특히, 정적 객체(나무, 건물)와 동적 객체(차량, 사람) 등 서로 다른 occlusion 객체 유형에 대한 분석과, 이미지 내 특정 위치(하단, 중심부, 배경)에 따른 마스킹 위치, 다양한 가림 면적 변화에 따른 모델의 강건성 변화를 추가적으로 수행하여 제안된 프레임워크의 실용성과 범용성을 더욱 심층적으로 검증할 계획이다.

REFERENCES

[1] J. Choi, K. C. Marsim, M. Jeong, K. Ryoo, J. Kim, and H. Myung, "Multi-unmanned aerial vehicle pose estimation based on visual-inertial-range sensor fusion," *Journal of Institute of Control, Robotics and Systems (in Korean)*, vol. 29, no. 11, pp. 859-865, Nov. 2023.

[2] J. Lee, H. J. Lee, S. K. Arachchige, N. Kim, H. Heo, and K. Lee, "Search operations with geolocation estimation of missing persons based on real-time drone images," *Journal of Institute of Control, Robotics and Systems (in Korean)*, vol. 30, no. 8, pp. 890-896, Aug. 2024.

[3] S. Hu, M. Feng, R. M. Nguyen, and G. H. Lee, "Cvm-net: Cross-view matching network for image-based ground-to-aerial geo-localization," *Proc. IEEE Conference on Computer Vision Pattern Recognition (CVPR)*, pp. 7258-7267, 2018.

[4] Y. Tian, C. Chen, and M. Shah, "Cross-view image matching for geo-localization in urban environments," *Proc. IEEE Conference on Computer Vision Pattern Recognition (CVPR),* pp. 3608-3616, 2017.

[5] S. Workman, R. Souvenir, and N. Jacobs, "Wide-area image geolocalization with aerial reference imagery," *Proc. IEEE Conference on Computer Vision Pattern Recognition (CVPR),* pp. 3961-3969, 2015.

[6] W. J. Ahn, S. Y. Park, D. S. Pae, H. D. Choi, and M. T. Lim, "Bridging viewpoints in cross-view geo-localization with Siamese vision transformer," *IEEE Transactions on Geoscience and Remote Sensing*, 2024.

[7] Y. Guo, M. Choi, K. Li, F. Boussaid, and M. Bennamoun, "Soft exemplar highlighting for cross-view image-based geo-localization," *IEEE Transactions on Image Processing*, vol. 31, pp. 2094–2105, 2022.

[8] G. Li, M. Qian, and G. S. Xia, "Unleashing unlabeled data: A paradigm for cross-view geo-localization," *Proc. IEEE/CVF Conference on Computer Vision Pattern Recognition (CVPR)*, pp. 16719–16729, 2024.

[9] Y. Shi, X. Yu, D. Campbell, and H. Li, "Where am I looking at? Joint location and orientation estimation by cross-view matching," *Proc. IEEE/CVF Conference on Computer Vision Pattern Recognition (CVPR)*, pp. 4064–4072, 2020.

[10] A. Toker, Q. Zhou, M. Maximov, and L. Leal-Taixé, "Coming down to earth: Satellite-to-street view synthesis for geo-localization," *Proc. IEEE/CVF Conference on Computer Vision Pattern Recognition (CVPR)*, pp. 6488–6497, 2021.

[11] H. Yang, X. Lu, and Y. Zhu, "Cross-view geo-localization with layer-to-layer transformer," *Advances in Neural Information Processing Systems*, vol. 34, pp. 29009–29020, 2021.

[12] X. Zhang, X. Li, W. Sultani, Y. Zhou, and S. Wshah, "Cross-view geo-localization via learning disentangled geometric layout correspondence," *Proc. The AAAI Conference on Artificial Intelligence*, vol. 37, no. 3, pp. 3480–3488, June 2023.

[13] J. Zhao, Q. Zhai, P. Zhao, R. Huang, and H. Cheng, "Co-visual pattern-augmented generative transformer learning for automobile geo-localization," *Remote Sensing*, vol. 15, no. 9, p. 2221, 2023.

[14] Y. Shi, L. Liu, X. Yu, and H. Li, "Spatial-aware feature aggregation for image based cross-view geo-localization," *Advances in Neural Information Processing Systems*, vol. 32, 2019.

[15] X. Wang, R. Xu, Z. Cui, Z. Wan, and Y. Zhang, "Fine-grained cross-view geo-localization using a correlation-aware homography estimator," *Advances in Neural Information Processing Systems*, vol. 36, pp. 5301–5319, 2023.

[16] S. Zhu, M. Shah, and C. Chen, "TransGeo: Transformer is all you need for cross-view image geo-localization," *Proc. IEEE/CVF Conference on Computer Vision Pattern Recognition (CVPR)*, pp. 1162–1171, 2022.

[17] S. Zhu, T. Yang, and C. Chen, "VIGOR: Cross-view image geo-localization beyond one-to-one retrieval," *Proc. IEEE/CVF Conference on Computer Vision Pattern Recognition (CVPR)*, pp. 3640–3649, 2021.

[18] Y. Zhu, H. Yang, Y. Lu, and Q. Huang, "Simple, effective and general: A new backbone for cross-view image geo-localization," arXiv preprint arXiv:2302.01572, 2023.

[19] S. Lee, C. Sung, J. Park, and H. Myung, "Enhancing cross-view geo-image retrieval using contrastive learning," *Proc. The 39th Institute of Control, Robotics, and Systems Annual Conference (in Korean)*, pp. 358–359, 2024.

[20] F. Deuser, K. Habel, and N. Oswald, "Sample4Geo: Hard negative sampling for cross-view geo-localisation," *Proc. IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 16847–16856, 2023.

[21] J. Park, C. Sung, S. Lee, D. Kang, and H. Myung, "Cross-view geo-localization via effective negative sampling," *Proc. The 24th International Conference on Control, Automation, and Systems (ICCAS)*, pp. 1078-1083, 2024.

[22] K. He, X. Chen, S. Xie, Y. Li, P. Dollár, and R. Girshick, "Masked autoencoders are scalable vision learners," *Proc. IEEE/CVF Conference on Computer Vision Pattern Recognition (CVPR)*, pp. 16000–16009, 2022.

[23] Z. Xie, Z. Zhang, Y. Cao, Y. Lin, J. Bao, Z. Yao, Q. Dai, and H. Hu, "SimMIM: A simple framework for masked image modeling," *Proc. IEEE/CVF Conference on Computer Vision Pattern Recognition (CVPR)*, pp. 9653–9663, 2022.

[24] H. Bao, L. Dong, S. Piao, and F. Wei, "BEiT: BERT pre-training of image transformers," arXiv preprint arXiv:2106.08254, 2021.

[25] Z. Liu, H. Mao, C.-Y. Wu, C. Feichtenhofer, T. Darrell, and S. Xie, "A ConvNet for the 2020s," *Proc. IEEE/CVF Conference on Computer Vision Pattern Recognition (CVPR)*, pp. 11976–11986, 2022.

[26] G. Ghiasi, Y. Cui, A. Srinivas, R. Qian, T.-Y. Lin, E. D. Cubuk, Q. V. Le, and B. Zoph, "Simple copy-paste is a strong data augmentation method for instance segmentation," *Proc. IEEE/CVF Conference on Computer Vision Pattern Recognition (CVPR)*, pp. 2918–2928, 2021.

이 승 희

2013년 이화여자대학교 전자공학과 (공학사). 2017년 KAIST 건설및환경공학과 (공학석사). 2017년~현재 KAIST 전기및전자공학부 박사과정. 2013~2015년 LG 전자 TV연구소, 2022~2023년 미국 John Deere. 2024년~현재 국방과학연구소 연구원. 관심분야는 로봇비전, Cross-view geo-image retrieval, 무인기자율화, 임무계획, 상황인식.