Evolving Visual Place Recognition

2025.01.24 겨울 세미나

Sogang University
Vision & Display Systems Lab, Dept. of Electronic Engineering

Presented By
김동규

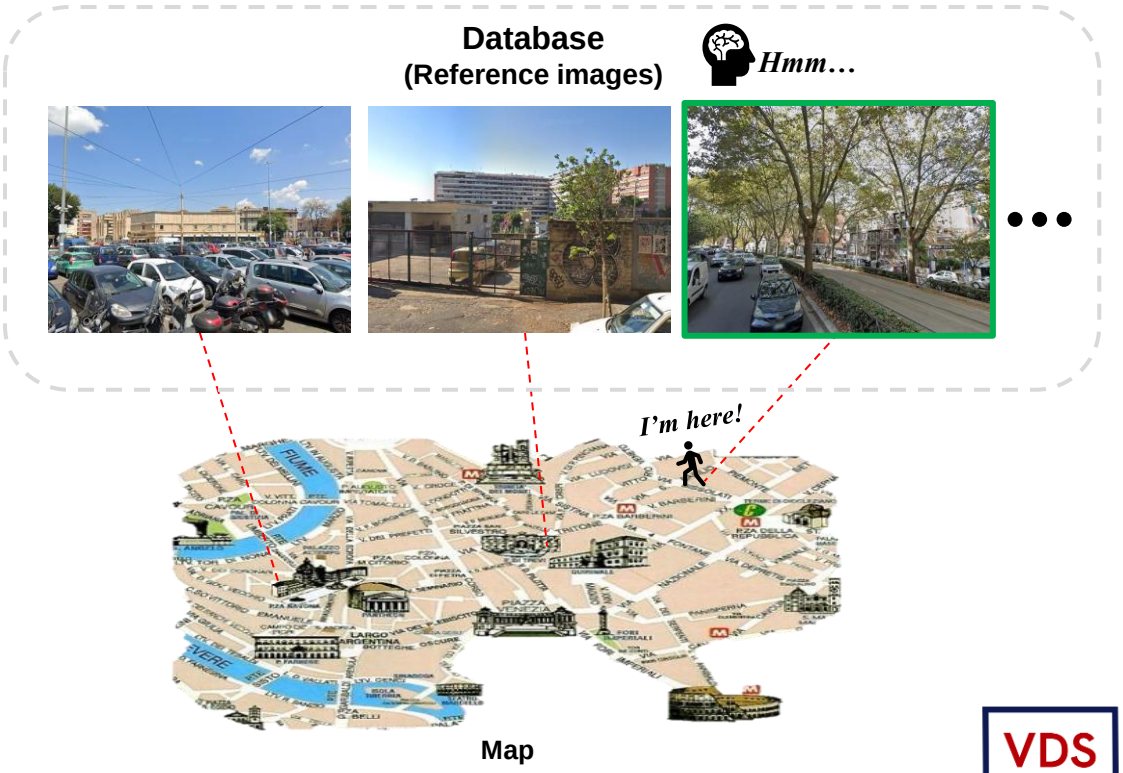
Outline

- Background
  - What is Visual Place Recognition(VPR)?
  - Advancements in VPR
  - Challenges in VPR

- Paper
  - BoQ: A Place is Worth a Bag of Learnable Queries (CVPR 2024)
  - Revisit Anything: Visual Place Recognition via Image Segment Retrieval (ECCV 2024)

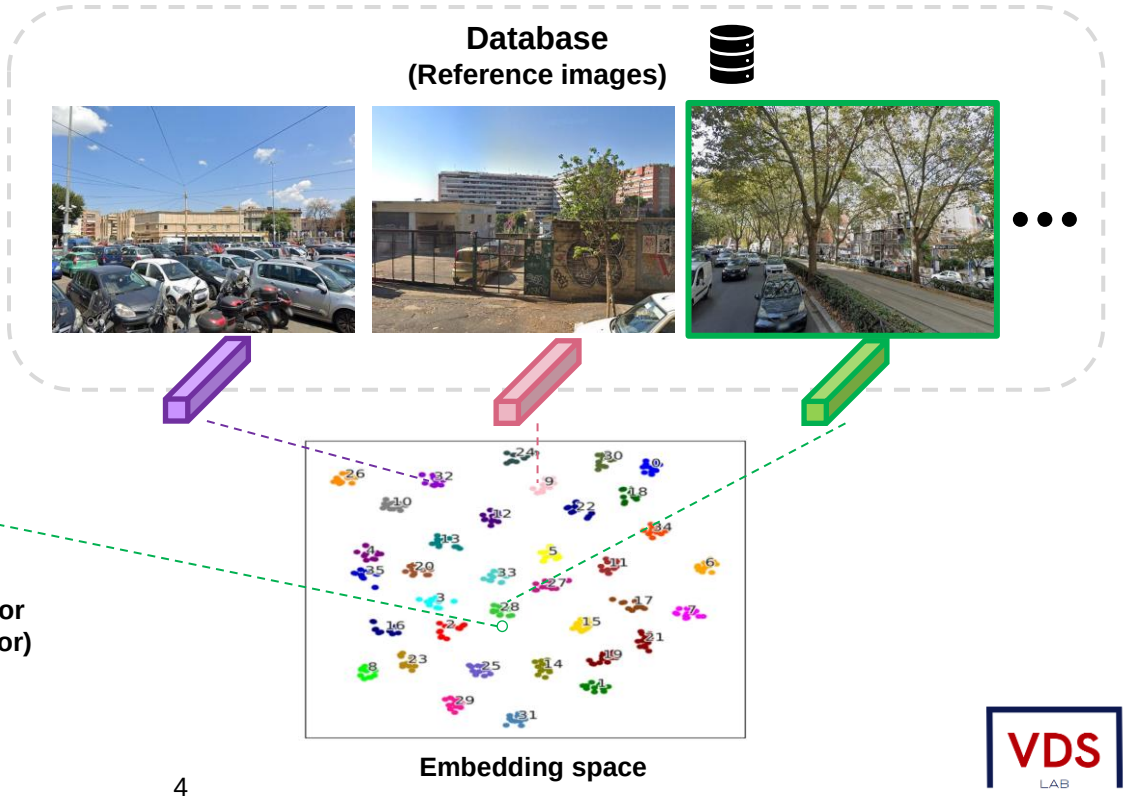
Background

- What is Visual Place Recognition (VPR)?
  - Image-based place recognition
  - 주어진 query image와 database 내의 reference images 를 매칭하여 현재 위치를 추정
    - Map과 해당 map의 images (database) 를 알고 있다는 전제가 필요
    - GPS 의존도가 낮음
    - 단일 카메라 사용
  - 적용 분야
    - Autonomous driving
    - SLAM



Background

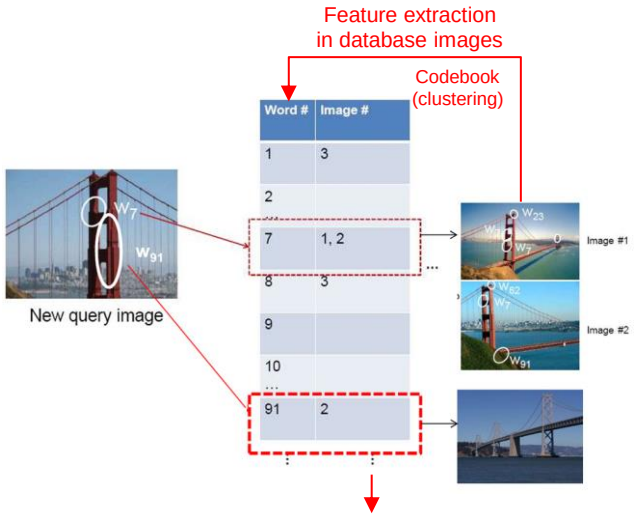
- What is Visual Place Recognition (VPR)?
  - 대규모 database를 기반으로 처리하기 때문에 메모리 문제가 존재
  - Image에서 중요한 정보를 추출, 결합하여 descriptor를 생성
    - Embedding space에서 같은 장소의 이미지는 가깝게, 먼 장소의 이미지는 멀게 학습
  - VPR model
    - Feature extraction
    - Aggregation
    - Matching
  - Image의 place를 잘 표현하는 descriptor를 만드는 model



1) Jégou, Hervé, et al. "Aggregating local descriptors into a compact image representation." *2010 IEEE computer society conference on computer vision and pattern recognition*. IEEE, 2010.

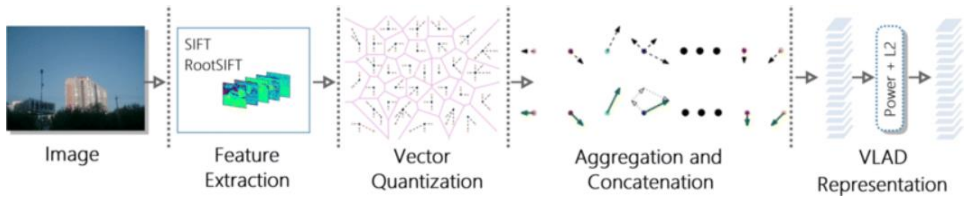
Background

- Advancements in VPR
  - 딥러닝 이전의 VPR: handcrafted feature-based method
    - Bag of Visual Words
      - Local descriptor들의 출현 빈도에 따라 recognition
        - ✓ Query image에서 visual words를 extraction
        - ✓ Local descriptor 를 count하여 histogram으로 표현
        - ✓ Histogram의 similarity를 이용하여 recognition



< BoVW의 histogram 생성과정 >

- VLAD (Vector of Locally Aggregated Descriptors) [1)]
  - Local descriptor와 centroid (중앙값)의 차이의 집합으로 모은 후 vector화
    - ✓BoVW와 동일하게 codebook을 만들고 matching
    - ✓각 local descriptor과 codebook간의 residual을 계산
    - ✓계산된 차이를 aggregation 하여 하나의 통합된 vector로 표현

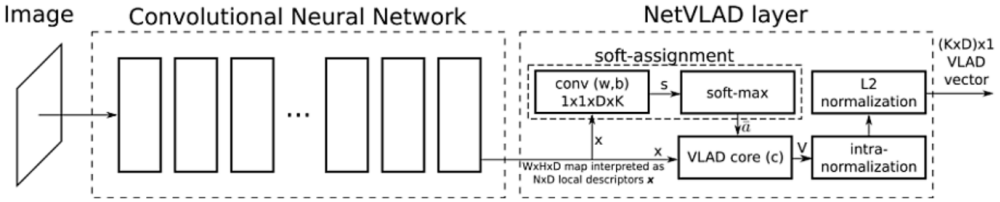


< VLAD의 descriptor생성과정 >

1) Arandjelovic, Relja, et al. "NetVLAD: CNN architecture for weakly supervised place recognition." *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016.
2) Lee, Chen-Yu, Patrick W. Gallagher, and Zhuowen Tu. "Generalizing pooling functions in convolutional neural networks: Mixed, gated, and tree." *Artificial intelligence and statistics. PMLR*, 2016.

Background

- Advancements in VPR
  - 딥러닝 이후의 VPR: deep learning-based method
    - NetVLAD[1)]
      - VLAD를 end-to-end 딥러닝 기반으로 확장
        - ✓ 각 local descriptor과 codebook간의 residual을 학습



< NetVLAD의 descriptor생성과정 >

- GeM (Generalized Mean) Pooling [2)]
  - Local descriptor의 공간적 정보를 잘 반영
    - ✓Global average pooling와 max pooling의 일반화 형태

p = 1 p = ∞

$p = 1$ $p = 3$ $p = 10$

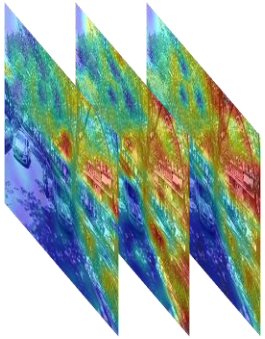
$$f_k^{(g)} = \left(\frac{1}{|\chi_k|} \sum_{x \in \chi_k} x^{p_k} \right)^{\frac{1}{p_k}}$$

p : learnable parameter

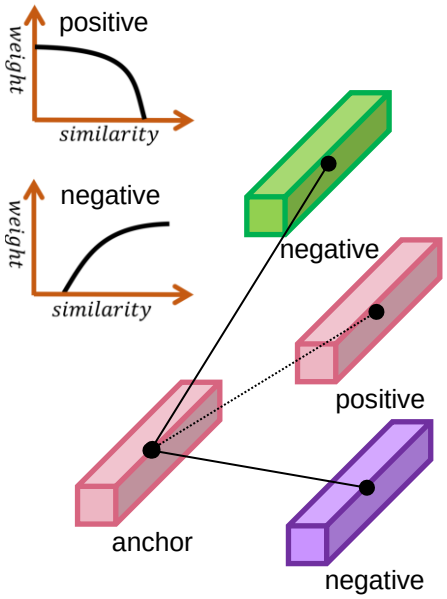
Background

- Advancements in VPR
  - Standard VPR pipeline

Image



feature



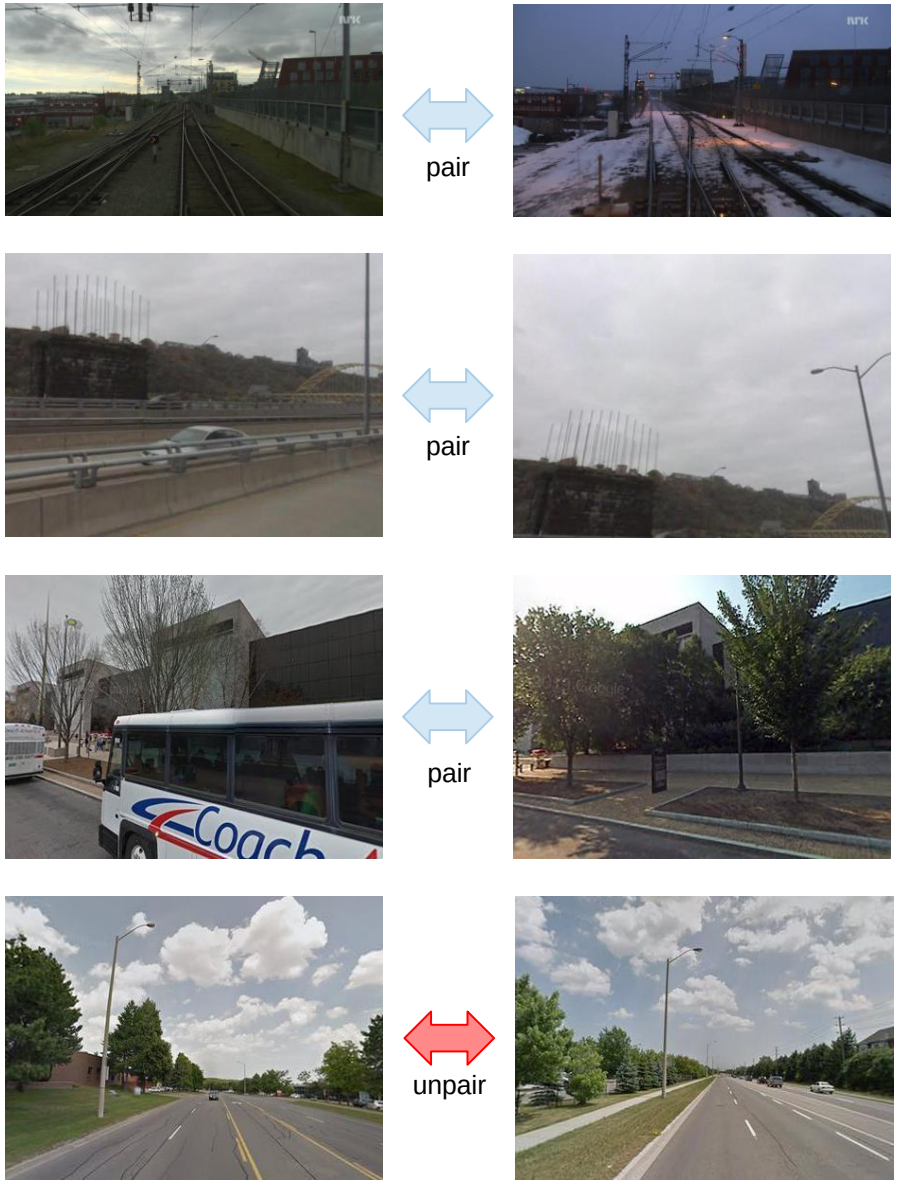
Global descriptor

Top-k images

Feature extraction	Aggregation	Loss	Re-ranking (optional)
• DINOv2 • ResNet • ConvNeXt • MobileNet • … • Fine-tunning • Parameter efficient fine-tunning	• BoQ • SALAD • MixVPR • NetVLAD • GeM • …	• Multi-similarity • Triplet • Quadruplet • Contrastive • …	• Superglue • RANSAC • …

Background

- Challenges in VPR
  - Memory efficiency
    - Large-scale data handling
    - Real-time processing
  - Robustness
    - Environmental variations
      - 날씨, 조명, 계절 등 환경적 요인
    - Viewpoint variations
      - 카메라 각도 변화
    - Noise and occlusion
      - 장애물로 인한 가려짐
    - Perceptual aliasing
      - 유사한 구조를 띄는 장소



< VPR의 challenge 예시 >

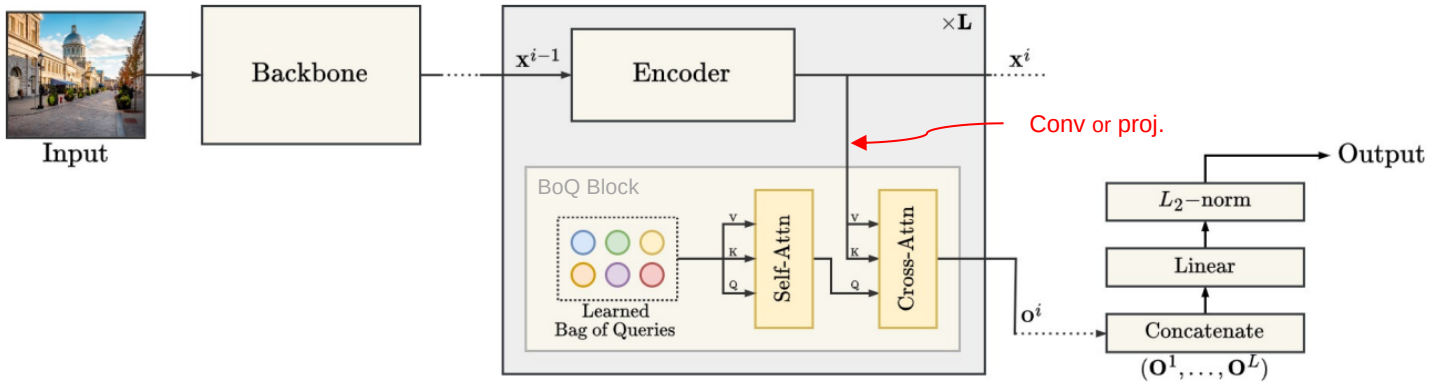
Ali-bey, Amar, et al. "BoQ: A Place is Worth a Bag of Learnable Queries." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024.

BoQ

- Motivation
  - 기존 aggregation 방법론의 한계
    - 기존 방법들은 local feature 를 aggregation 하여 global descriptor를 생성
      - 해당 과정에서 local descriptor들을 효과적으로 aggregation 하지 못하는 경우가 존재
      - 또는, 성능 개선을 위해 re-ranking을 방법을 도입하여 시간과 메모리를 많이 소모
  - 효율적인 aggregation을 통한 global descriptor 생성 기술의 필요성
    - BoQ(Bag of Queries) aggregation 방법 도입
      - Global query 를 사용하여 local feature들을 cross-attention으로 탐색하고 aggregation
      - CNN, ViT backbone에 모두 통합 사용 가능

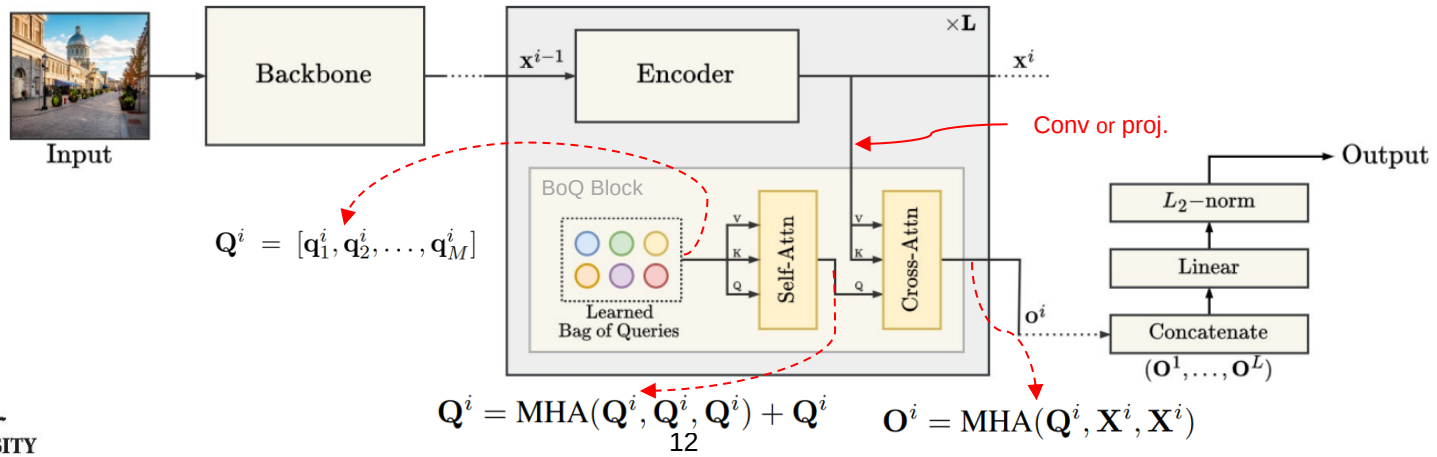
BoQ

- Methodology
  - Feature extraction
    - Backbone 에서 여러 layer에서의 다양한 high-level feature 를 추출
      - 일반적으로 backbone 의 초반 layer는 low-level, 후반 layer는 high-level을 추출
        - ✓ 이러한 계층적 정보를 활용하여 더 풍부하고 강력한 descriptor를 생성
      - CNN-based backbone
        - ✓ 각 layer에서 얻은 feature를 3x3 convolution을 통해 차원을 조정
      - ViT-based backbone
        - ✓ 각 layer에서 얻은 patch 단위의 feature를 linear projection을 통해 차원 조정



BoQ

- Methodology
  - Aggregation (BoQ block)
    - Learnable global queries
      - 각 query는 input feature에서 특정 정보를 탐색하고 aggregation 하도록 학습
    - Self-attention
      - Query간의 관계를 학습하고, input feature와 상관 없이 query를 refine
    - Cross-attention
      - Refine 된 global queries는 input feature와 cross-attention을 통해 정보를 통합
    - Output은 다른 BoQ block의 output과 통합을 거쳐 global descriptor를 생성



1) Ali-bey, Amar, Brahim Chaib-draa, and Philippe Giguere. "Gsv-cities: Toward appropriate supervised visual place recognition." *Neurocomputing 513*.2022.
2) Wang, Xun, et al. "Multi-similarity loss with general pair weighting for deep metric learning." *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2019.

BoQ

- Experiments
  - Datasets
    - Training set: Google Street View (GSV-cities)[1)]
    - Test set:

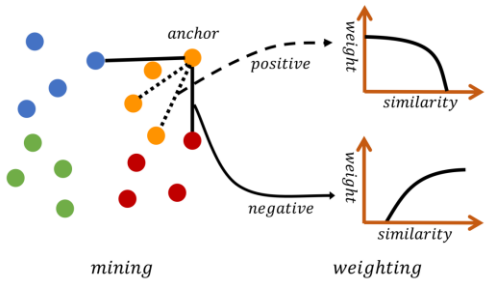
Dataset name	# quer.	# ref.	Variations		
			Viewpoint	Season	Day/Night
MSLS [50]	740	18.9k	×	××	×
Pitts250k [44]	8.2k	84k	×××		
Pitts30k [44]	6.8k	10k	×××		
AmsterTime [51]	1231	1231	×××	××	×
Eynsham [16]	24k	24k	×		
Nordland* [53]	2760	27.6k		×××	
Nordland** [42]	27.6k	27.6k		×××	
St-Lucia [34]	1464	1549			
SVOX [10]	14.3k	17.2k	×	×××	×××
SPED [53]	607	607		××	××

< GSV dataset 예시 >


< AmsterTime dataset 예시 >

  - Loss
    - Multi-similarity loss[2)]

$$\mathcal{L}_{MS} = \frac{1}{B}\sum_{q=1}^{B}\left\{\frac{1}{\alpha}\log\left[1+\sum_{p\in\mathcal{P}_q} e^{-\alpha(S_{qp}-\lambda)}\right] + \frac{1}{\beta}\log\left[1+\sum_{n\in\mathcal{N}_q} e^{\beta(S_{qn}-\lambda)}\right]\right\},$$



BoQ

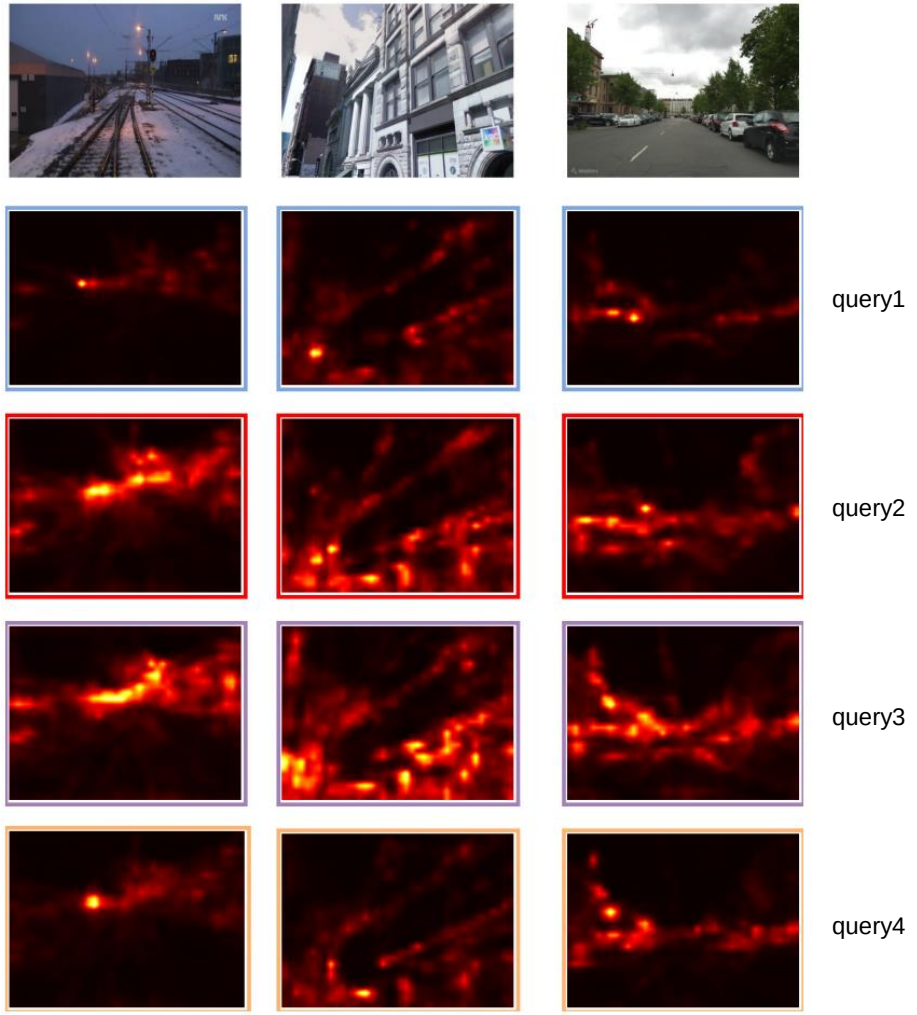
- Experiments

Method	Dim.	Pitts250k-test			MSLS-val			SPED			Nordland*		
		R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10
AVG [6]	2048	78.3	89.8	92.6	73.5	83.9	85.8	58.8	77.3	82.7	15.3	27.4	33.9
GeM [39]	2048	82.9	92.1	94.3	76.5	85.7	88.2	64.6	79.4	83.5	20.8	33.3	40.0
NetVLAD [6]	32768	90.5	96.2	97.4	82.6	89.6	92.0	78.7	88.3	91.4	32.6	47.1	53.3
SPE-NetVLAD [52]	163840	89.2	95.3	97.0	78.2	86.8	88.8	73.1	85.5	88.7	25.5	40.1	46.1
Gated NetVLAD [55]	32768	89.7	95.9	97.1	82.0	88.9	91.4	75.6	87.1	90.8	34.4	50.4	57.7
Conv-AP [3]	4096	92.4	97.4	98.4	83.4	90.5	92.3	80.1	90.3	93.6	38.2	54.8	61.2
CosPlace [11]	2048	92.3	97.4	98.4	87.4	93.8	94.9	75.3	85.9	88.6	54.4	69.8	75.9
MixVPR [4]	4096	94.2	98.2	98.9	88.0	92.7	94.6	85.2	92.1	94.6	58.4	74.6	80.0
EigenPlaces [12]	2048	94.1	97.9	98.7	89.2	93.6	95.0	82.4	91.4	94.7	54.2	68.0	73.9
BoQ (Ours)	4096	**95.0**	**98.4**	**99.1**	91.1	**94.8**	95.7	85.4	93.1	95.4	69.5	83.4	87.0
BoQ (Ours)	16384	**95.0**	98.3	99.0	**91.4**	94.5	**96.1**	**86.2**	**94.4**	**96.1**	**74.4**	**86.1**	**89.8**

Method	Backbone	Multi-view datasets			Frontal-view datasets							
		AmsterTime	Eynsham	Pitts30k	Nordland**	St Lucia	SVOX Night	SVOX Overcast	SVOX Rain	SVOX Snow	SVOX Sun
NetVLAD [6]	VGG-16	16.3	77.7	85.0	13.1	64.6	8.0	66.4	51.5	54.4	35.4
SFRS [20]	VGG-16	29.7	72.3	89.1	16.0	75.9	28.6	81.1	69.7	76.0	54.8
CosPlace [11]	VGG-16	38.7	88.3	88.4	58.5	95.3	44.8	88.5	85.2	89.0	67.3
EigenPlaces [12]	VGG-16	38.0	89.4	89.7	54.5	95.4	42.3	89.4	83.5	89.2	69.7
Conv-AP [3]	ResNet-50	33.9	87.5	90.5	62.9	99.7	43.4	91.9	82.8	91.0	80.4
CosPlace [11]	ResNet-50	47.7	90.0	90.9	71.9	99.6	50.7	92.2	87.0	92.0	78.5
MixVPR [4]	ResNet-50	40.2	89.4	91.5	76.2	99.6	64.4	96.2	91.5	96.8	84.8
EigenPlaces [12]	ResNet-50	48.9	90.7	**92.5**	71.2	99.6	58.9	93.1	90.0	93.1	86.4
BoQ (Ours)	ResNet-50	**53.0**	**91.5**	92.4	**85.5**	**99.9**	**85.2**	**98.3**	**96.4**	**98.4**	**96.5**

BoQ

- Experiments
  - Learnable queries의 특성 시각화
  - 64개중 4개의 queries 시각화
    - 각 query는 다른 영역에 집중
    - 특정 패턴이나 local 정보 강조
      - 일부는 특정 물체
      - 일부는 배경 패턴
    - BoQ가 다양한 정보를 통합함을 보임



< 학습된 query별 시각화 >

BoQ

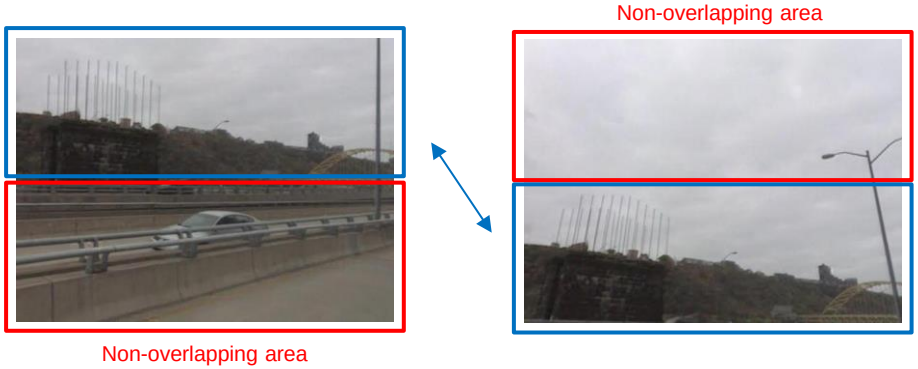
- Conclusion
  - 기존의 aggregation 방법보다 효율적이고 강건한 global descriptor를 생성
    - Learnable queries와 cross-attention으로 중요한 정보를 동적으로 aggregation
  - 다양한 VPR dataset에서 높은 성능을 보임
    - 특히 계절 변화와 시점 변화와 같은 challenge한 환경에서도 강함
  - CNN 및 ViT backbone 모두에 적용 가능하며 우수한 성능을 보임
    - 다양한 작업에서 global representation 생성을 위해 사용가능
    - VPR 외의 다른 vision task에도 적용 가능성을 가짐
- Future works
  - 성능을 확장하여 더 큰 dataset 복잡한 환경 그리고 실시간 응용에서 활용 가능성
  - 더 가벼운 구조로 개선하여 계산 비용을 줄이는 방향으로 고려

Garg, Kartik, et al. "Revisit Anything: Visual Place Recognition via Image Segment Retrieval." European Conference on Computer Vision (ECCV), 2024.

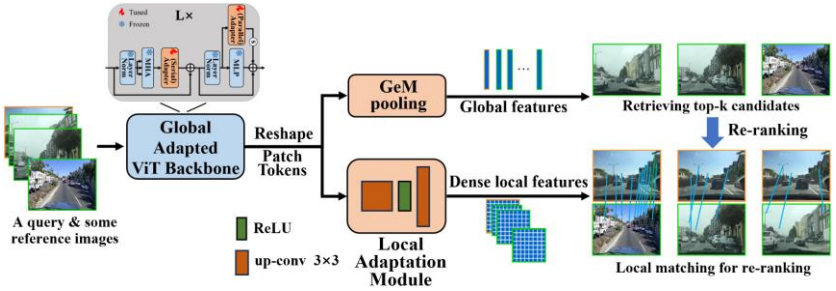
1) Lu, Feng, et al. "Towards seamless adaptation of pre-trained models for visual place recognition." *The Twelfth International Conference on Learning Representations* (2024).

Revisit Anything

- Motivation
  - 기존 VPR 방법론의 문제점
    - 기존 VPR은 전체 image를 global descriptor로 변환하여 matching
      - 이는 view-point가 변화할 때 특히 취약함
      - Non-overlapping 구역이 matching 성능에 부정적인 영향을 끼침
    - 기존 VPR에서 Local descriptor의 사용 방식
      - local descriptor를 aggregation하여 global descriptor를 만드는 용도
      - Re-ranking 용도



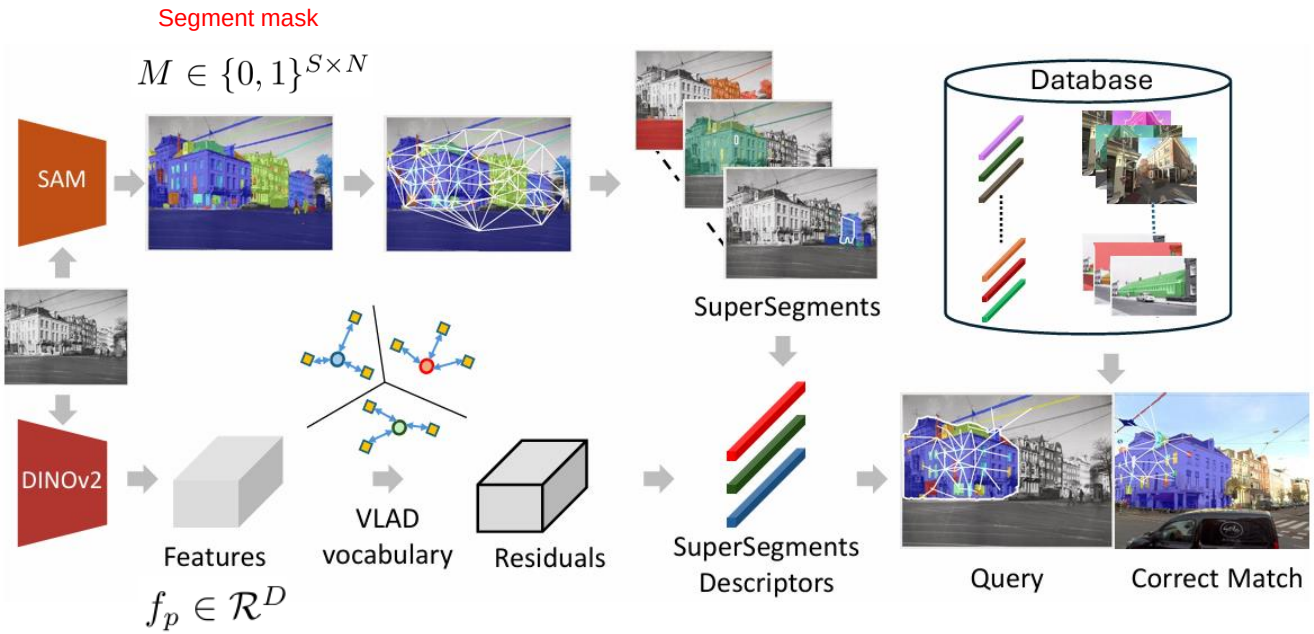
< View-point 차이가 큰 data 예시 >



< 기존 VPR model인 SelaVPR[1]의 예시 >

Revisit Anything

- Methodology
  - Problem formulation
    - 기존 VPR 방법들은 global descriptor를 사용하여 image를 전체적으로 표현
      - 따라서 view-point 의 변화로 인해 겹치는 부분이 적을 경우, matching 성능이 저하
    - 이를 해결하기 위해 image를 **segment descriptor의 집합**으로 표현
      - Segment level의 representation은 부분적인 정보를 효과적으로 표현



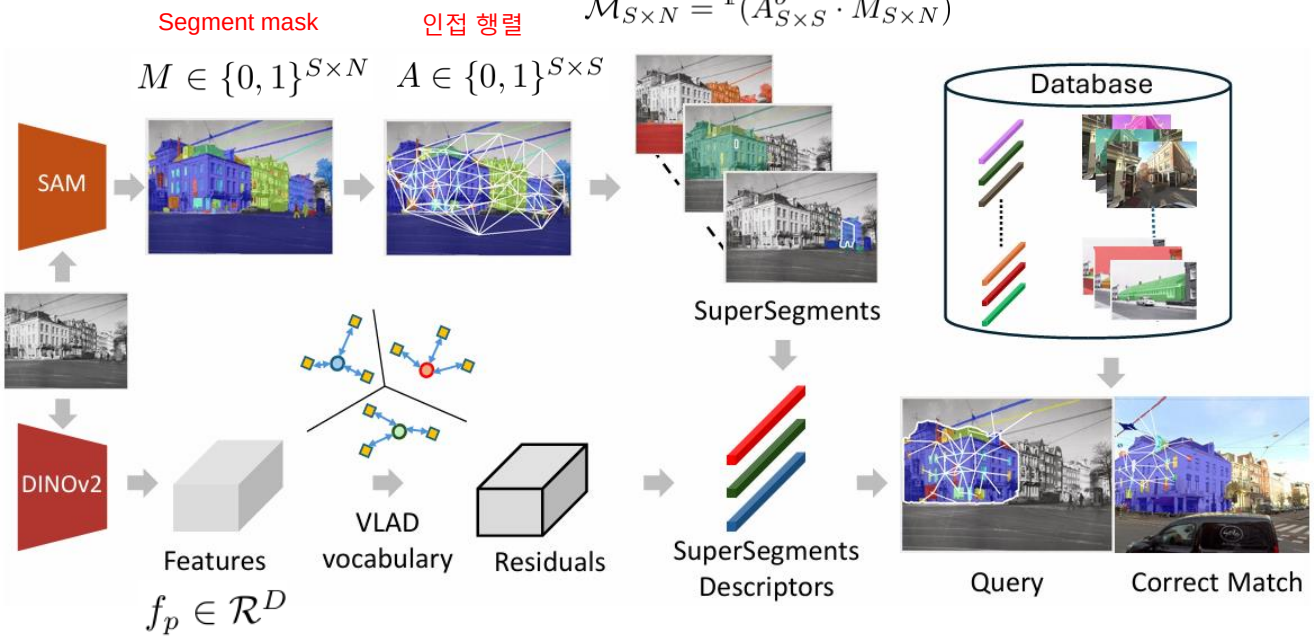
Revisit Anything

- Methodology
  - SuperSegment
    - 단순한 segment를 넘어 각 segment와 그 주변 이웃관계를 포함하는 확장된 segment
      - **Delaunay Triangulation**을 통해 세그먼트의 이웃 관계를 구성
        - ✓ 각 segment의 무게중심을 중심점으로 지정하고 인접 행렬 생성
        - ✓ 인접 행렬에서 **Neighborhood expansion**으로 각 segment에서 SuperSegment를 생성

< 각 점(좌)에서 Delaunay Triangulation이 적용된 경우(중)와 적용되지 않은 경우 (우) >

o= order

$$\mathcal{M}_{S\times N} = \mathbb{1}(A^{o}_{S\times S} \cdot M_{S\times N})$$



Revisit Anything

- Methodology
  - SuperSegment
    - Neighborhood expansion

- 동일 image의 여러 Supersegment – 중복 영역 존재

- 기존 segment와 다르게 풍부한 representation와 함께 맥락 정보를 함께 제공
- View-point 변화와 occlusion에 강인함

Revisit Anything

- Methodology
  - SuperSegment Descriptors
    - Segments, Neighborhood segments, image 모두에 적용가능한 통합 descriptor 제안

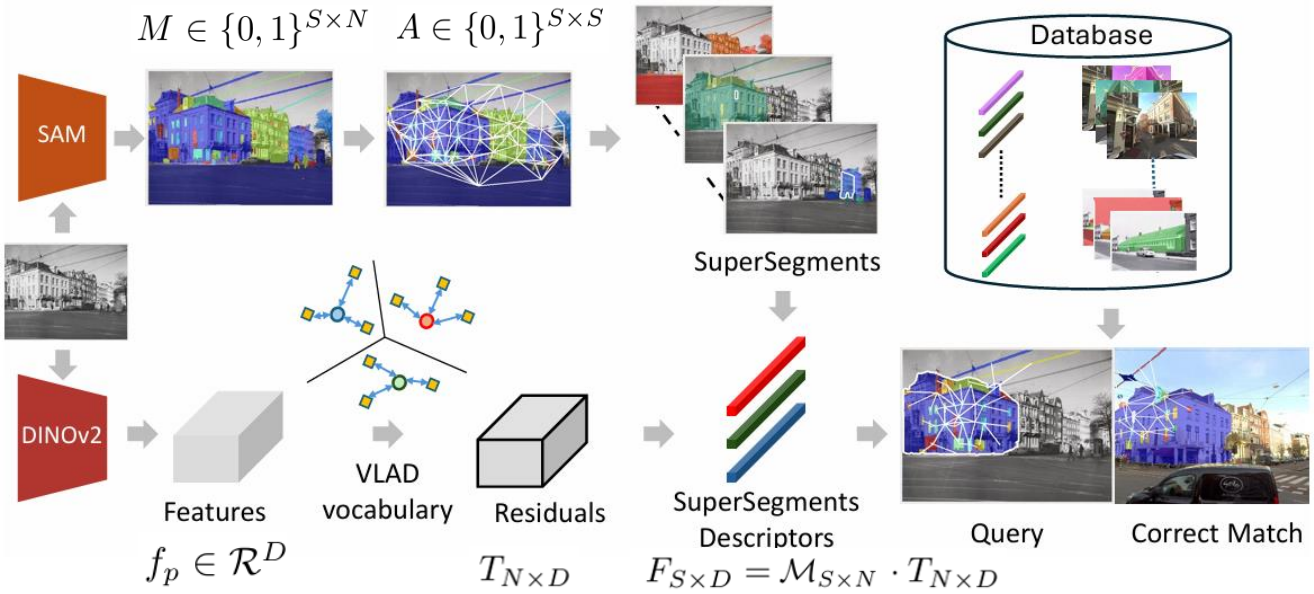
SuperSegment masks Aggregated features (VLAD)

SuperSegment descriptors $F_{S\times D} = \mathcal{M}_{S\times N} \cdot T_{N\times D}$

- Segment mask의 개수에 따라 다양한 방법 사용가능
  - ✓ Ex) N = 1 인 경우, image 전체

SuperSegment masks

Segment mask 인접 행렬 $\mathcal{M}_{S\times N} = {}^{\mathbb{1}}(A^{o}_{S\times S} \cdot M_{S\times N})$



Pixel feature Pixel descriptor

Revisit Anything

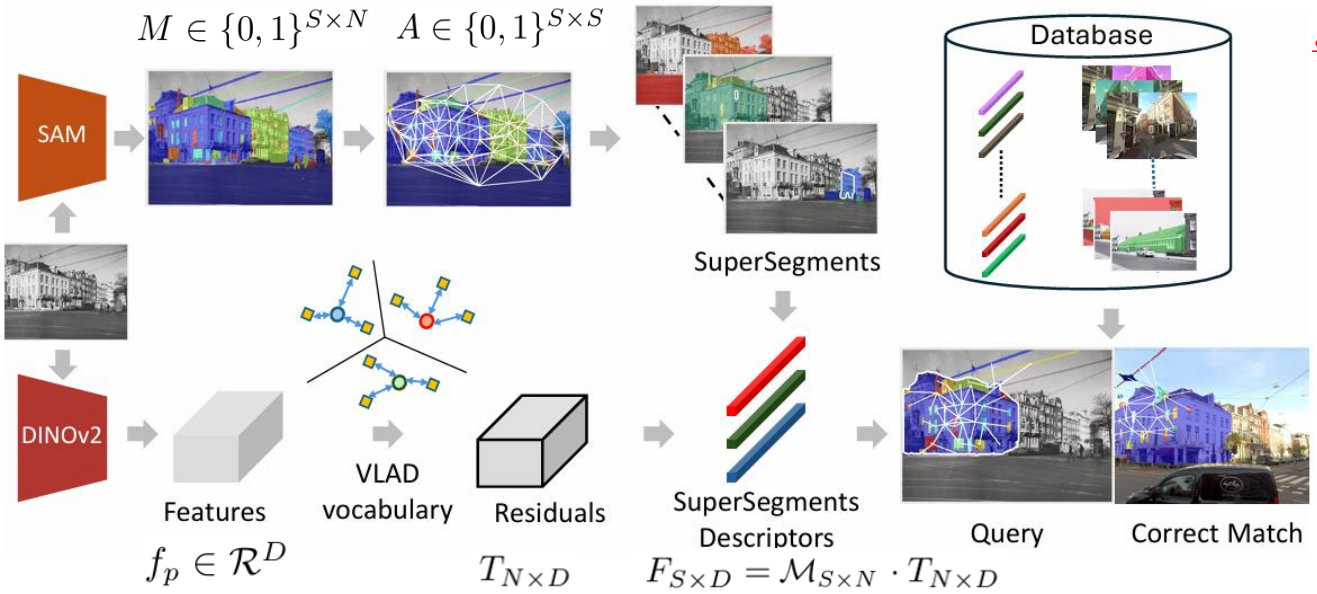
- Methodology
  - Image Retrieval via Segments
    - Query image의 각 SuperSegment descriptors 이용해 database의 모든 SuperSegment와 비교
      - 찾은 SuperSegment descriptor가 속한 image들을 검색 하여 가장 유사한 image 찾기
      - 이때, 매칭되는 descriptor들의 weights의 합을 기준으로 판단

$$r_j^* = \underset{r_j}{\operatorname{argmax}}\ \hat{\theta}(r_j)$$

$$\hat{\theta}(r_j) = \sum_{s=1}^{S}\sum_{k=1}^{K'} \theta_{sk} \cdot \mathbb{1}_{\{r_{sk}=r_j\}}$$

유사도 상위 K개 descriptor

Segment mask 인접 행렬 $\mathcal{M}_{S\times N} = {}^{\mathbb{1}}(A^o_{S\times S} \cdot M_{S\times N})$



Pixel feature

Pixel descriptor

Revisit Anything

- Experiments
  - Outdoor dataset

Method	Pitts-30K	MSLS SF	MSLS CPH	SF-XL Val	RO5k Med	RO5k Hard	RP6k Med	RP6k Hard
CosPlace	90.4/95.7	93.4/**97.5**	84.9/92.0	94.6/97.6	85.7/87.1	27.1/45.7	94.3/95.7	7.1/15.7
MixVPR	91.5/95.5	91.3/95.9	87.1/92.4	87.8/93.8	68.6/80.0	32.9/54.3	94.3/100	10.0/32.9
EigenPlaces	92.6/96.7	92.6/97.1	87.1/92.8	**96.4/98.2**	85.7/88.6	42.8/57.1	95.7/98.6	4.3/11.4
AnyLoc	87.7/94.7	83.4/94.6	79.9/89.1	84.4/91.9	88.6/92.9	40.0/58.6	97.1/100	11.4/44.3
SALAD	92.6/96.5	91.7/97.1	**92.3/96.1**	93.6/97.3	82.9/90.0	37.1/54.3	95.7/98.6	**14.3/58.6**
SegVLAD-PreT	86.7/94.2	88.4/94.2	81.7/90.7	90.9/96.4	**91.4/95.7**	**60.0/81.4**	94.3/**100**	8.6/48.6
SegVLAD-FineT	**93.2/96.8**	**94.6**/97.1	90.9/95.7	94.9/98.1	87.1/**95.7**	51.4/70.0	95.7/**100**	10.0/48.6

DINOv2 + NetVLAD 로 backbone만 finetunning

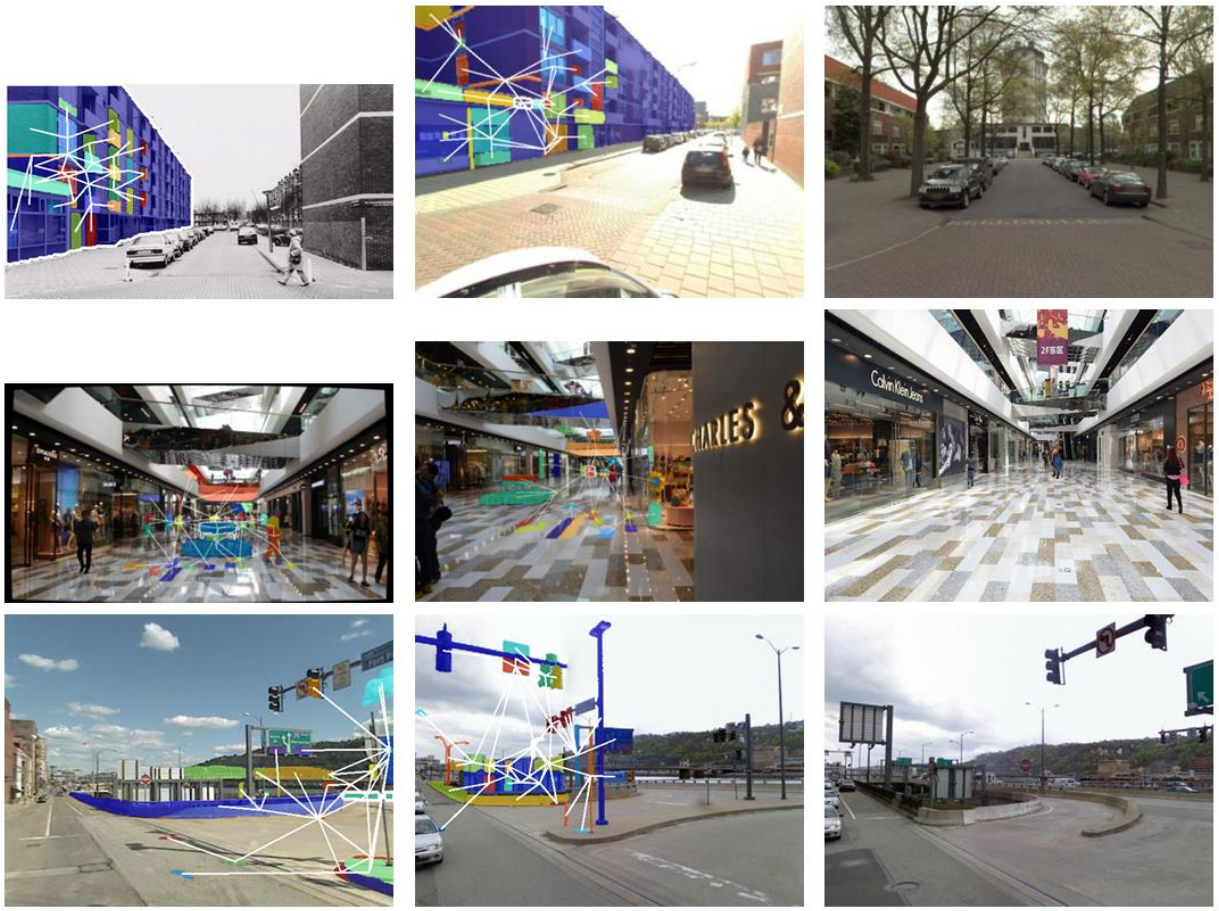
  - Out-of-distribution dataset
    - Fine-tunning 시 특정 domain에 overfitting 될 수 있음

	indoor	historical	Indoor-to-outdoor	indoor	aerial
Method	**Baidu**	**AmsterTime**	**InsideOut**	**17Places**	**VPAir**
CosPlace	41.6/55.0	47.7/69.8	0.2/2.0	81.3/88.2	4.6/13.7
MixVPR	64.4/80.3	40.2/59.1	0.0/1.8	85.2/90.1	6.8/16.1
EigenPlaces	56.5/72.8	48.9/69.5	0.4/1.4	83.0/90.1	6.5/17.9
AnyLoc	75.2/87.6	50.3/73.0	2.4/8.0	**95.3**/97.3	66.7/79.2
SALAD	74.8/86.5	55.4/75.6	0.6/1.8	82.5/88.2	25.8/38.7
SegVLAD-PreT	**78.5/93.8**	56.8/77.7	4.2/9.4	**95.3/98.0**	**69.8/83.7**
SegVLAD-FineT	68.1/89.0	**58.9/79.3**	**7.4/15.6**	95.1/97.5	35.4/55.3

Revisit Anything

- Experiments
  - 전역적인 배경보다 object에 집중하여 인간의 시각과 비슷하게 공간적인 맥락 파악

Query image | SegVLAD (ours) | AnyLoc[1)]



Revisit Anything

- Conclusion
  - Image segment 기반으로 새로운 VRP 방법인 SegVLAD 제안
    - 특정 instance를 인식하는 object retrieval과 유사
  - 기존의 global descriptor 기반의 model들보다 view-point 변화에 강함
  - 기존 연구와의 차별성 및 기존 연구의 패러다임 전환
- Future works
  - Segment 기반의 descriptor를 강건하게 구성
  - CLIP과 같은 LLM 을 통해 text기반 모델과의 통합 및 확장 가능성 제시
- Limitations
  - 많은 database를 위한 용량이 필요

감사합니다