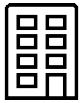
『2025 하계 세미나』

Visual Place Recognition



Pusan National University
Dept. of Electronics Engineering
Computer Vision & Signal Processing Lab.



Presented By
우민수

Outline

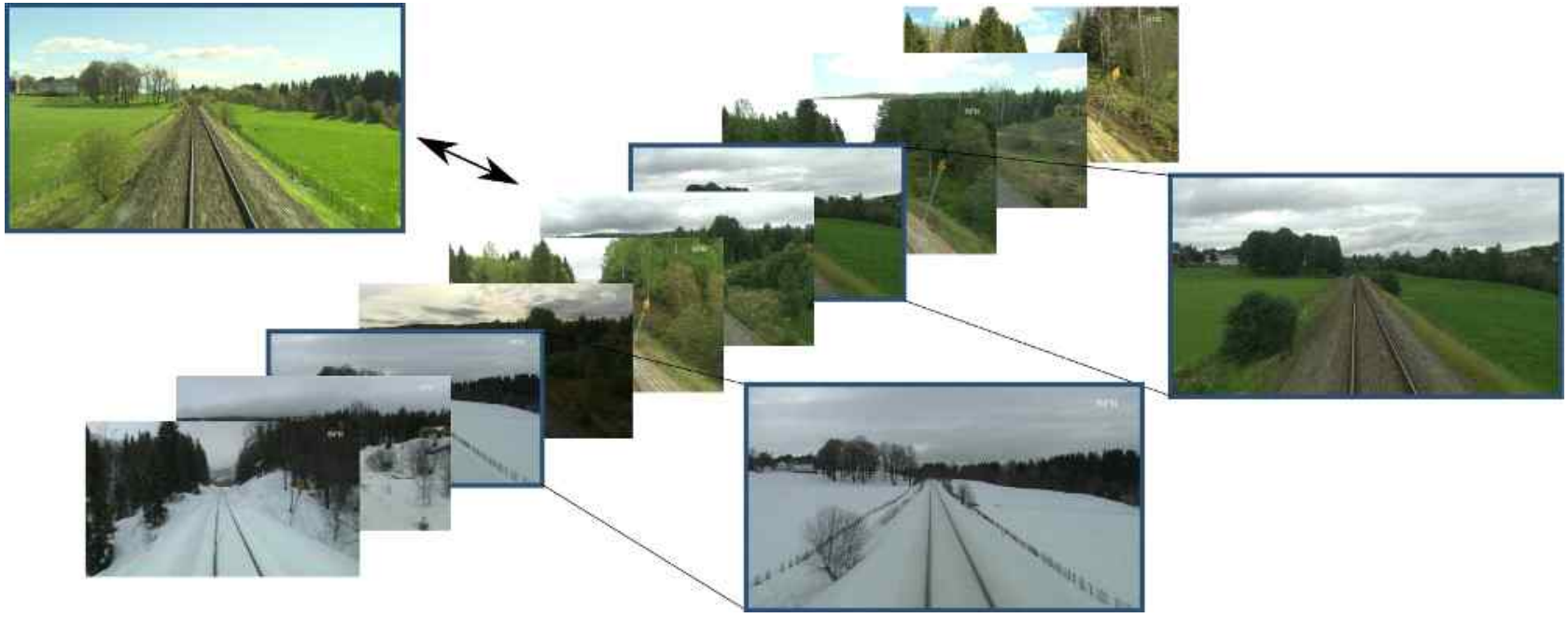
- 논문 선택 및 발표하는 이유
- Background
- ETR: An Efficient Transformer for Re-ranking in Visual Place Recognition [WACV 2023]
- R2Former: Unified Retrieval and Reranking Transformer for Place Recognition[CVPR 2023]

논문 선택 및 발표하는 이유

- LG CTO
  - Topic : Domain recognition
    - 로봇이 현재 어느 domain에 위치했는지 알고싶음
      - Domain recognition의 방법론으로써 Visual Place Recognition이 사용될수있음

Background

- Visual Place Recognition
  - 정의
    - 쿼리 이미지로 들어온 이미지와 데이터 베이스내의 가장 유사한 장소 이미지를 찾아내는 것

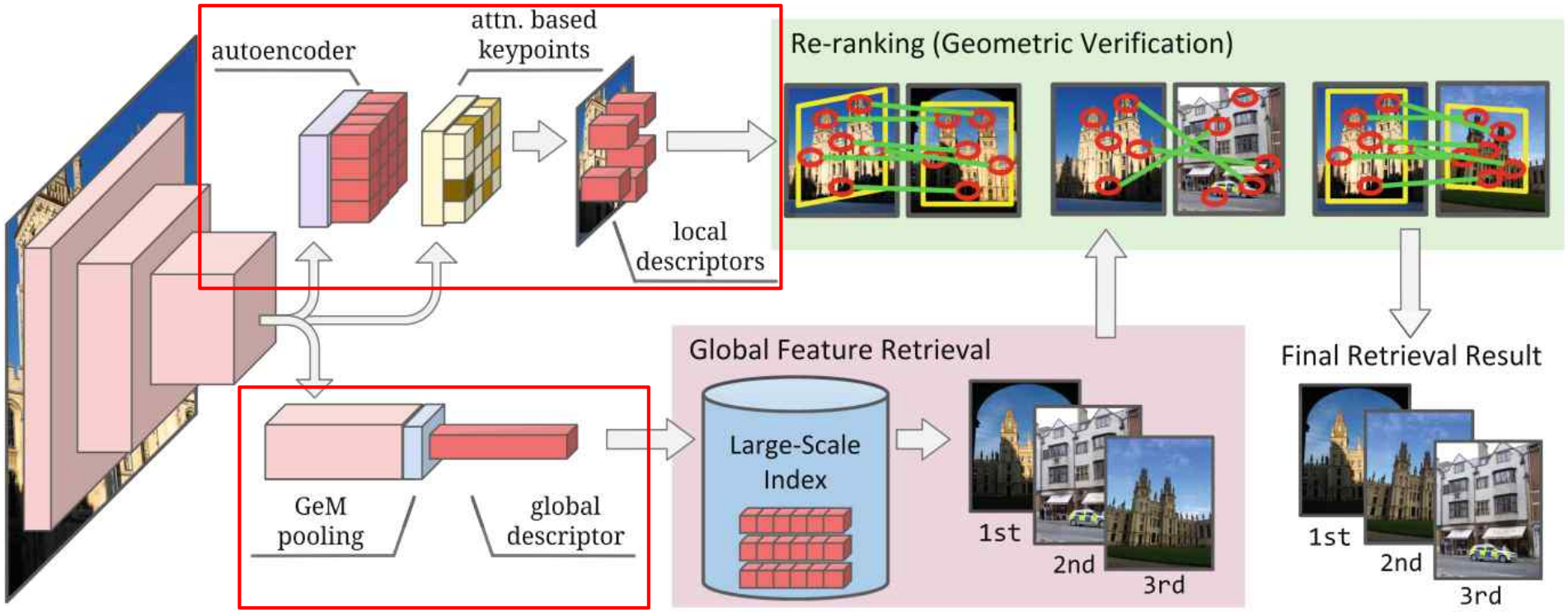


Background

- Global Descriptor
  - 정의
    - 이미지 전체를 단일 벡터로 요약하여 표현한 것
  - 특징
    - 이미지 전체의 전반적 색상, 질감, 구조와 같은 큰 스케일의 정보를 포함
    - 이미지를 하나의 고정된 벡터로 표현-> 빠른검색에 유리, 유사도 계산시 빠르게 동작
- Local Descriptor
  - 정의
    - 이미지의 특정 부분(패치 등)에서 추출된 feature 벡터 집합
  - 특징
    - 이미지 내의 지역적 패턴(물체의 모서리, 랜드마크 등)을 상세히 표현
    - 벡터가 다수 존재하므로(벡터 집합) 연산량이 크고 속도가 느림

Background

- DELG



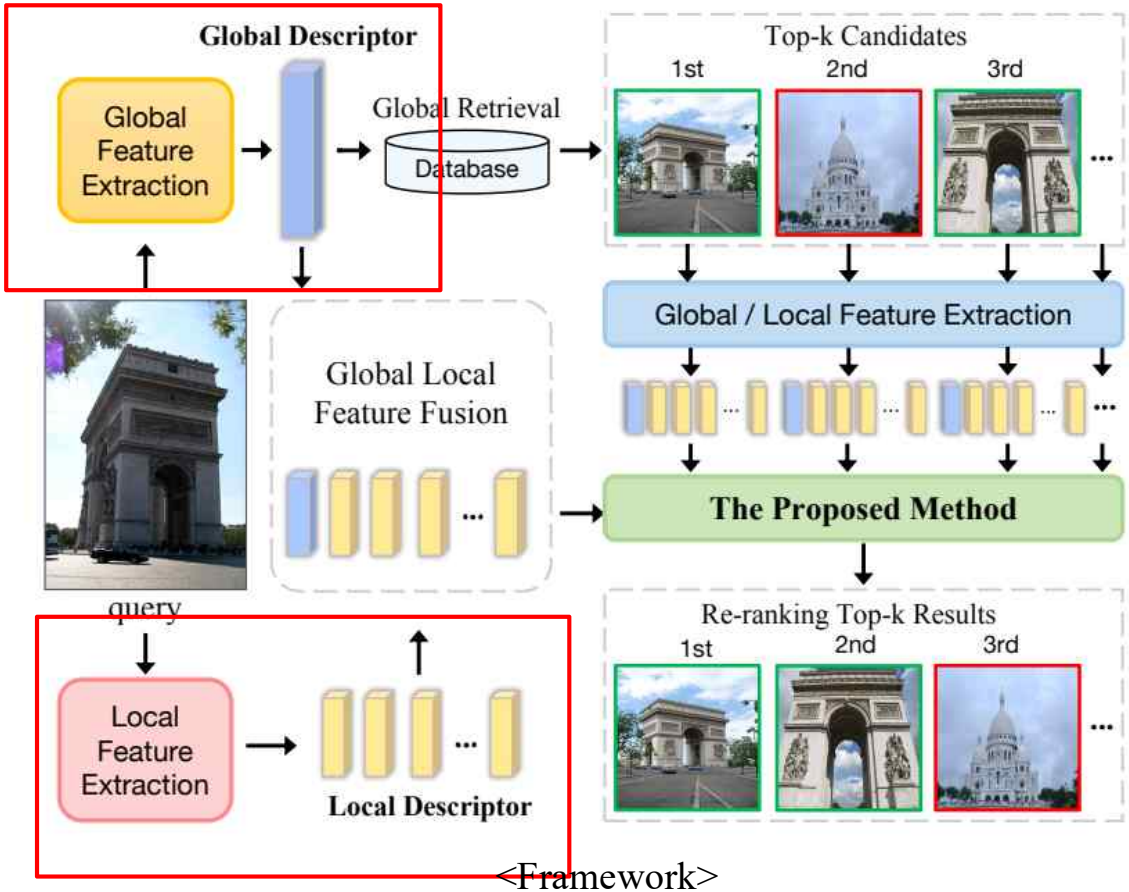
ETR: An Efficient Transformer for Re-ranking in Visual Place Recognition [WACV 2023]

Method

- Goal
  - Reranking에서 정확성과 속도의 균형을 달성
    - 기존의 방법들은 Local Descriptor를 기반으로 Reranking을 진행하는데 정확하지만 매우 느리고 계산량이 많음
    - ETR은 Transformer의 Self-Attention + Cross-Attention 구조를 활용해 Global·Local Descriptor를 동시에 처리하며 빠른 속도로 정확한 유사도 점수를 계산

Method

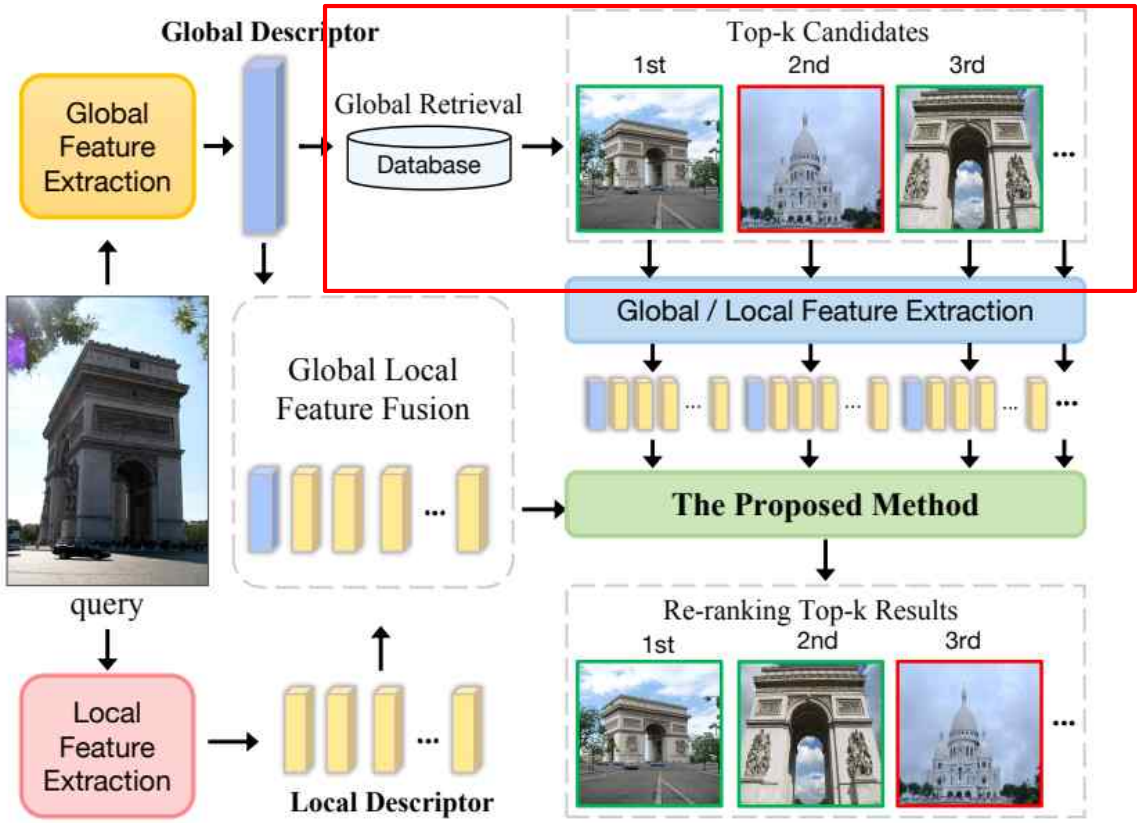
- Feature Extraction
  - Extract Global Descriptor, Local Descriptor
    - DELG(Resnet-50 기반 feature extractor)를 활용하여 쿼리 이미지에서 Global·Local Descriptor 추출



<Framework>

Method

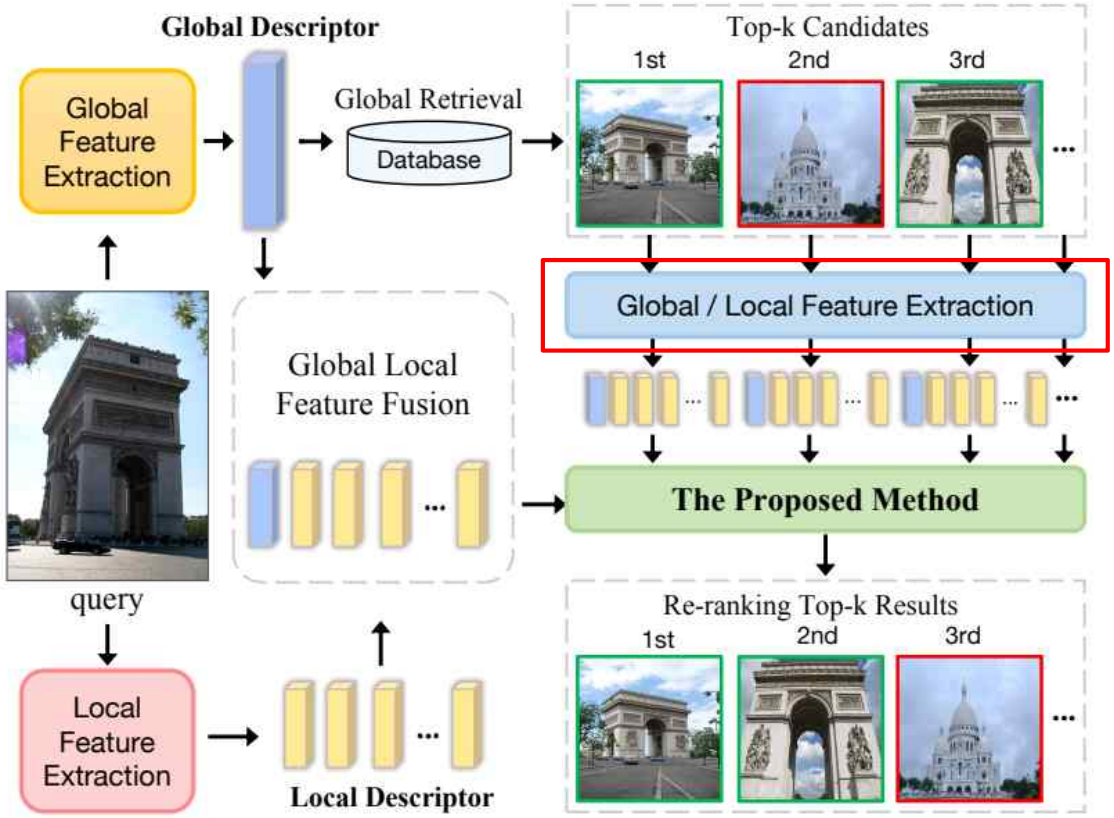
- Global Retrieval
  - Extract Top-k candidates
    - Global Descriptor 기반으로 유사도 측정하여 Top-k 이미지 획득



<Framework>

Method

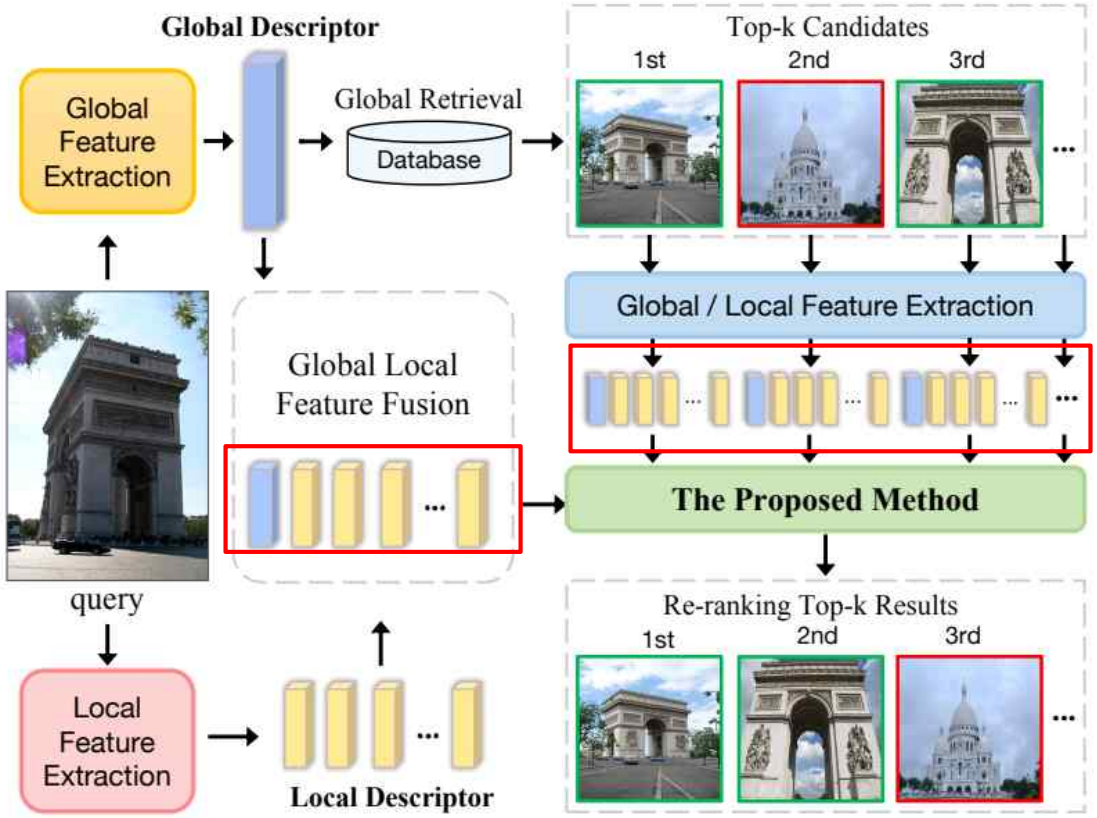
- Feature Extraction
  - Extract Global Descriptor, Local Descriptor
    - Top-k 이미지에 대해서 DELG(Resnet-50 기반 feature extractor)를 활용하여 Global·Local Descriptor 추출



<Framework>

Method

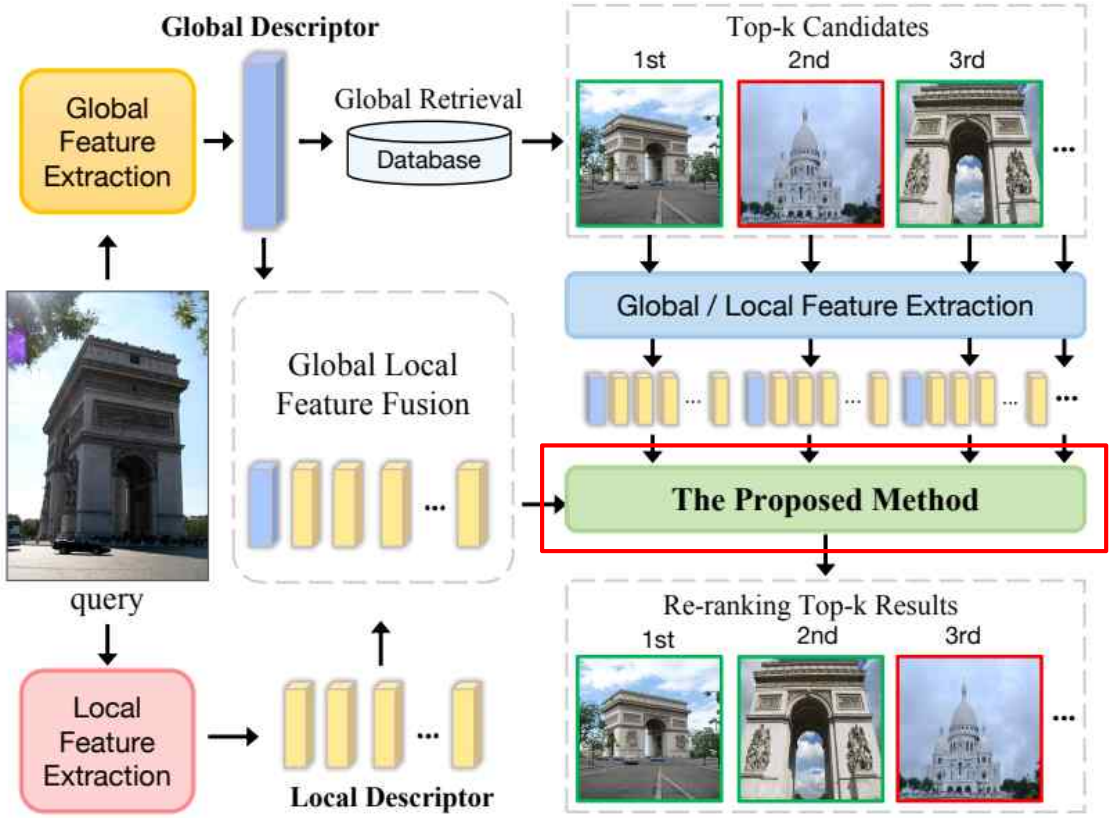
- Global Local Feature Fusion
  - Fusion Global Descriptor, Local Descriptor
    - Global·Local Descriptor 하나의 시퀀스로 합치기



<Framework>

Method

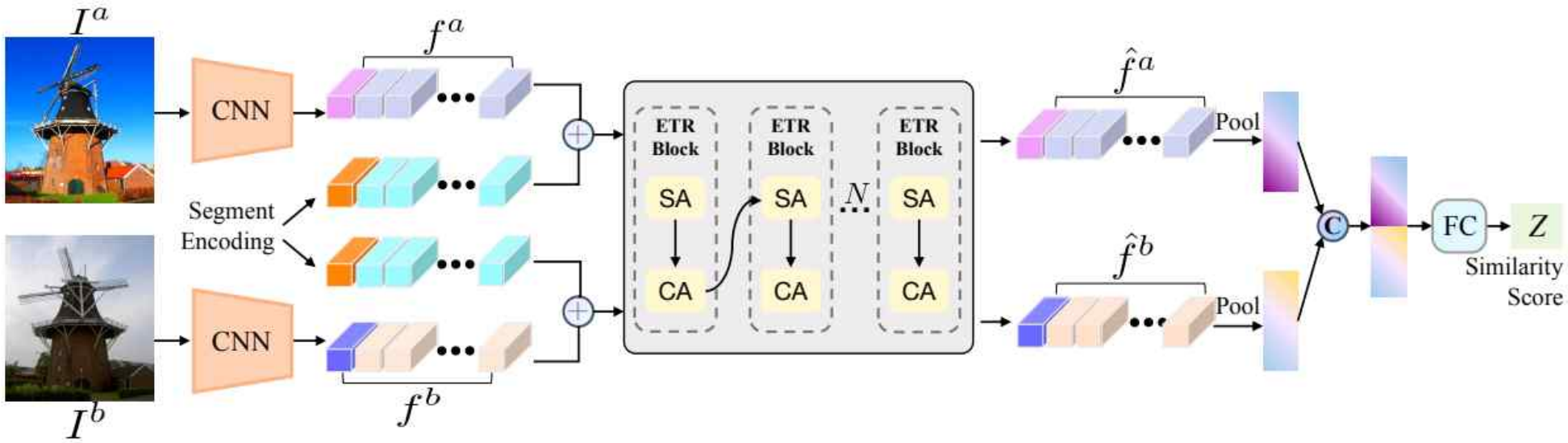
- The Proposed Method
  - Use ETR Block
    - 합쳐진 Global·Local Descriptor를 ETR Block에 넣어 유사도 재계산



<Framework>

Method

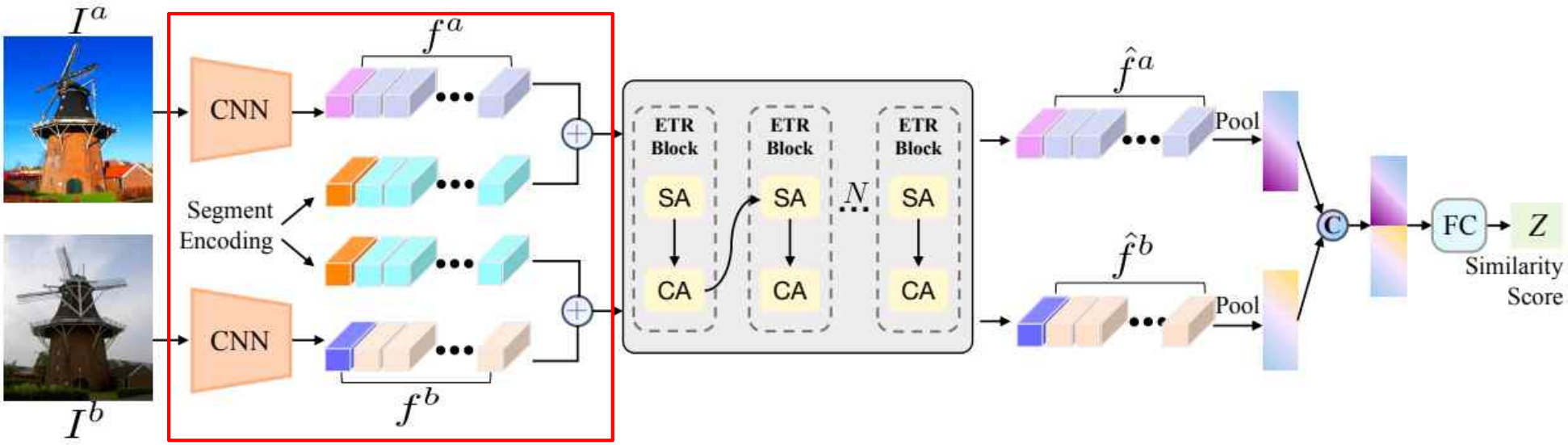
- The Proposed Method
  - Use ETR Block and Reranking
    - 합쳐진 Global·Local Descriptor를 ETR Block에 넣어 유사도 재계산



<Framework of Proposed Method>

Method

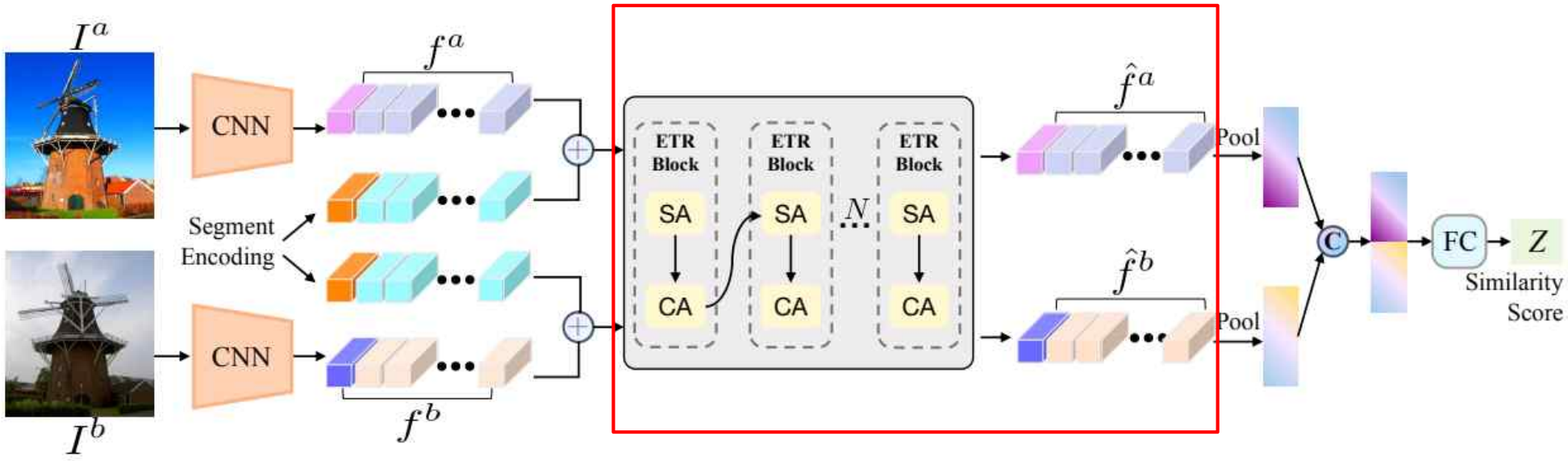
- The Proposed Method
  - Concat features with Segment Encoding
    - 합쳐진 Global·Local Descriptor중 Global Descriptor의 위치를 알기 위해서 Segment Encoding과 합쳐진 Global·Local Descriptor를 Concat



<Framework of Proposed Method>

Method

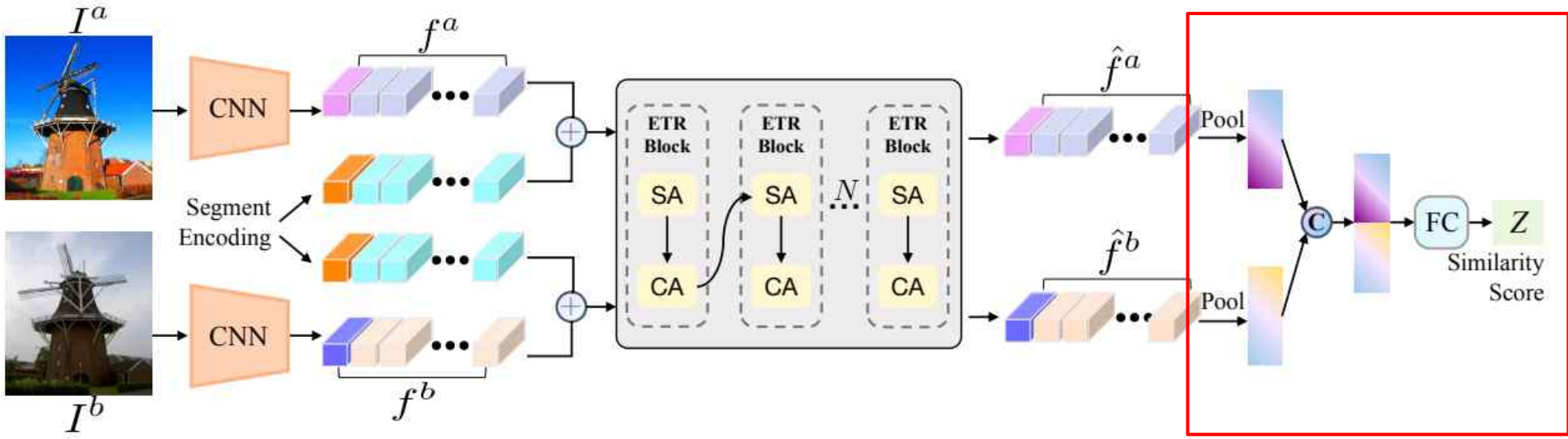
- The Proposed Method
  - Use ETR Block
    - Self-attention과 Cross-attention으로 이뤄진 ETR Block을 사용하여 합쳐진 Global·Local Descriptor가 서로 비교되고 보완됨



<Framework of Proposed Method>

Method

- The Proposed Method
  - Calculate Similartiy Score
    - 보완된 Global·Local Descriptor는 average pooling을 이용해 Flatten됨 이후 FC Layer를 지나 새로운 Similarity Score를 계산하는데 사용됨

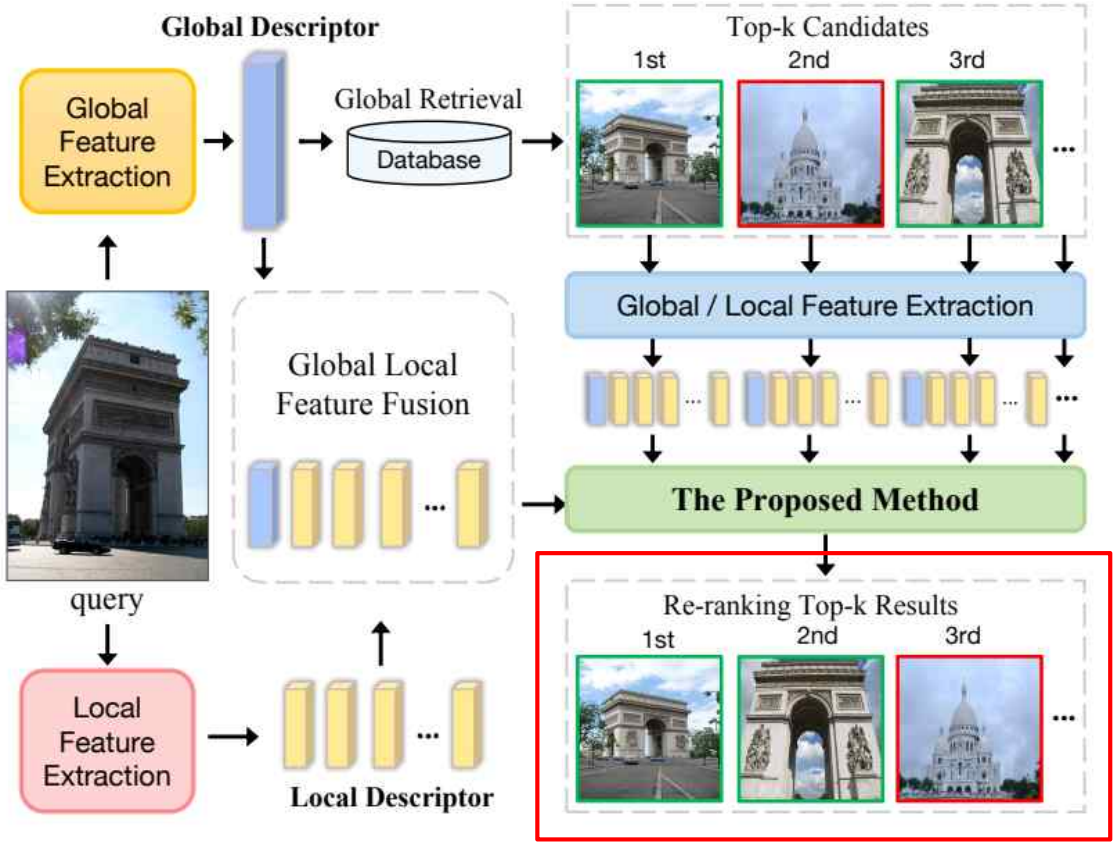


<Framework of Proposed Method>

부산대학교 PUSAN NATIONAL UNIVERSITY
CVSP Lab Computer Vision & Signal Processing

Method

- The Proposed Method
  - Reranking
    - 재계산된 유사도로 Top-k 이미지 Reranking



<Framework>

부산대학교 PUSAN NATIONAL UNIVERSITY
CVSP Lab Computer Vision & Signal Processing

Method

- Training and Loss function
  - Transformer 기반의 ETR Block을 학습
    - ETR블록을 통해 구한 Similarity Score[0,1]와 정답(1)과 오답(0)을 비교하여 Binary Cross-Entropy (BCE) Loss를 통해 ETR 블록내의 파라미터를 학습

$$L = -\frac{1}{N}\sum_{i=1}^{N}[y_i \log(\sigma(z_i)) + (1-y_i)\log(1-\sigma(z_i))]$$

- z_i: 쿼리 – 후보 이미지 쌍의 유사도 점수
- $y_i \in \{0,1\}$:레이블
  - $y = 1$:쿼리와 후보 이미지가 동일한 장소
  - $y = 0$:서로 다른 장소
- $\sigma(z) = \frac{1}{1+e^{-z}}$

<Loss 수식>

Experiment

- Training Dataset
  - v2-clean split
    - 관광지, 랜드마크 이미지가 라벨과 함께 제공되는 대규모 벤치마크 데이터셋
    - 322008장의 이미지 사용

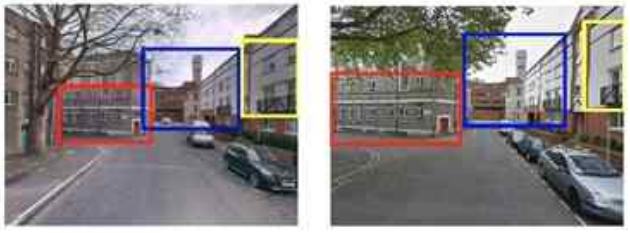


Experiment

- Test Dataset
  - MSLS
    - 160만개의 거리 이미지 포함 데이터셋
    - 총 1900개의 쿼리이미지 사용



  - Pitts30k
    - 도시 환경에서 촬영된 이미지로 건물, 도로등이 주요 요소
    - 총 6816개의 쿼리 이미지 사용



  - Tokyo 24/7
    - 도쿄에서 스마트폰으로 촬용된 이미지, 조명 변화가 심함
    - 총 315개의 쿼리 이미지 사용



Experiment

- Metric
  - Recall@N
    - 쿼리 이미지
      - 쿼리 이미지의 상위 N개 검색 결과 중 하나라도 Ground Truth와 일치하면 정답

Experiment

- Quantiative Results

Method	MSLS val			Pitts30k			Tokyo 24/7			MSLS challenge		
	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10
SFRS [15]	69.2	80.3	83.1	89.4	94.7	95.9	85.4	91.1	93.3	42.5	53.7	58.0
NetVLAD [1]	58.6	71.2	76.1	81.9	91.2	93.7	67.0	77.8	80.3	35.1	47.4	51.7
AP-GEM [29]	65.0	75.7	78.2	80.7	91.4	94.0	58.4	69.5	74.3	30.2	41.3	47.1
DELG global [6]	72.2	81.4	84.6	78.4	87.4	91.6	73.0	83.5	87.0	39.3	52.6	58.2
RRT [35]	72.4	86.5	89.0	80.7	90.7	93.9	86.7	94.0	94.9	39.1	55.4	63.0
SP-SuperGlue [10, 31]	78.4	82.8	84.2	87.2	94.8	96.4	88.2	90.2	90.2	50.6	56.9	58.3
DELG local [6]	**83.2**	**89.3**	89.5	**89.8**	**95.3**	**96.6**	86.4	92.4	93.0	52.2	61.9	65.4
Patch-NetVLAD-s [16]	77.8	84.3	86.5	87.5	94.5	94.8	70.2	78.7	82.2	48.1	59.4	62.3
Patch-NetVLAD-p [16]	79.5	86.2	87.7	88.7	94.5	95.9	86.0	88.6	90.5	48.1	57.6	60.5
ETR-S (ours)	80.5	86.5	88.9	83.1	91.1	93.8	**90.1**	93.0	94.6	**53.9**	**62.8**	**66.1**
ETR-D (ours)	79.3	88.0	**89.6**	84.2	91.6	93.8	89.2	**94.3**	**95.2**	50.6	62.1	65.8

<Quantiative Results>

Experiment

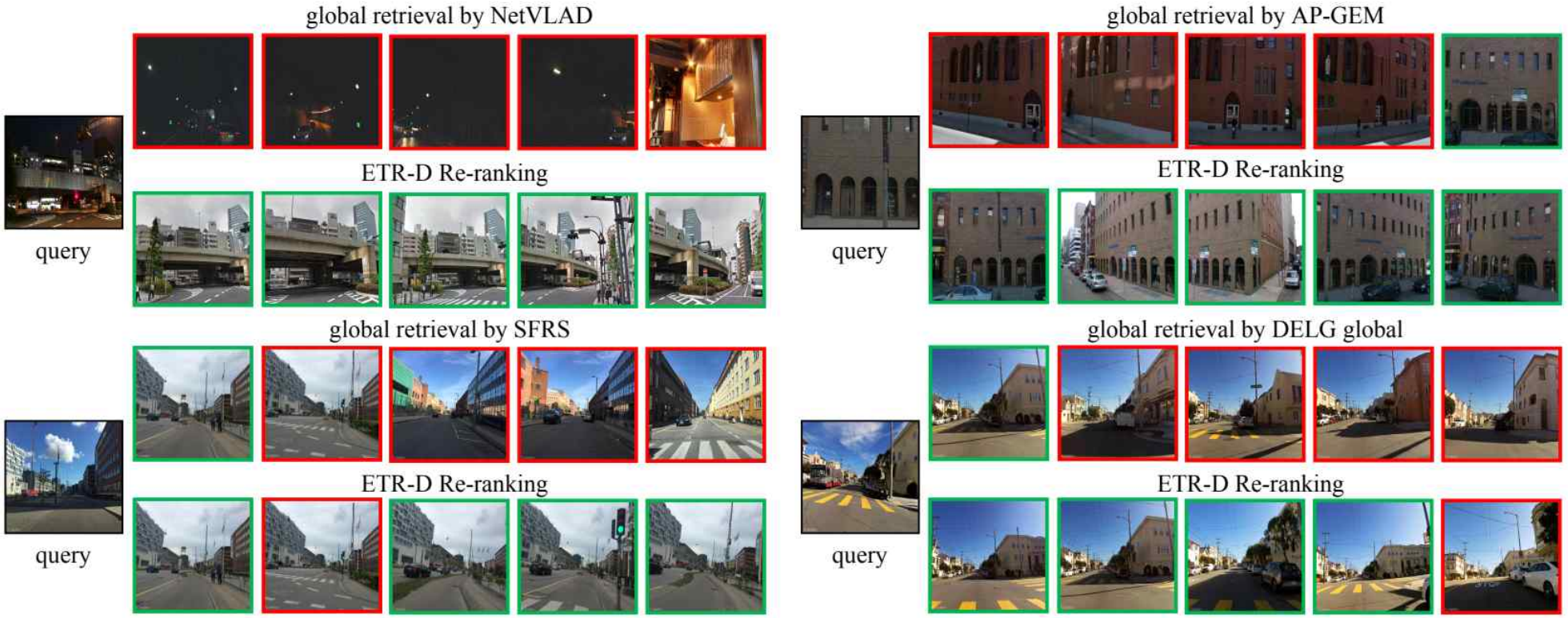
- Quantiative Results

Method	Extraction latency (ms)	Re-ranking latency (ms)	Memory (MB)
DELG local [6]	152	73100	0.9
Patch-NetVLAD-s [16]	21	200	1.9
Patch-NetVLAD-p [16]	487	7700	44.2
SP-SuperGlue [10, 31]	7	3800	0.7
RRT [35]	152	8	0.5
ETR-D (ours)	152	14	0.5
ETR-S (ours)	7	6	0.7

<Quantiative Results>

Experiment

- Qualitative Results



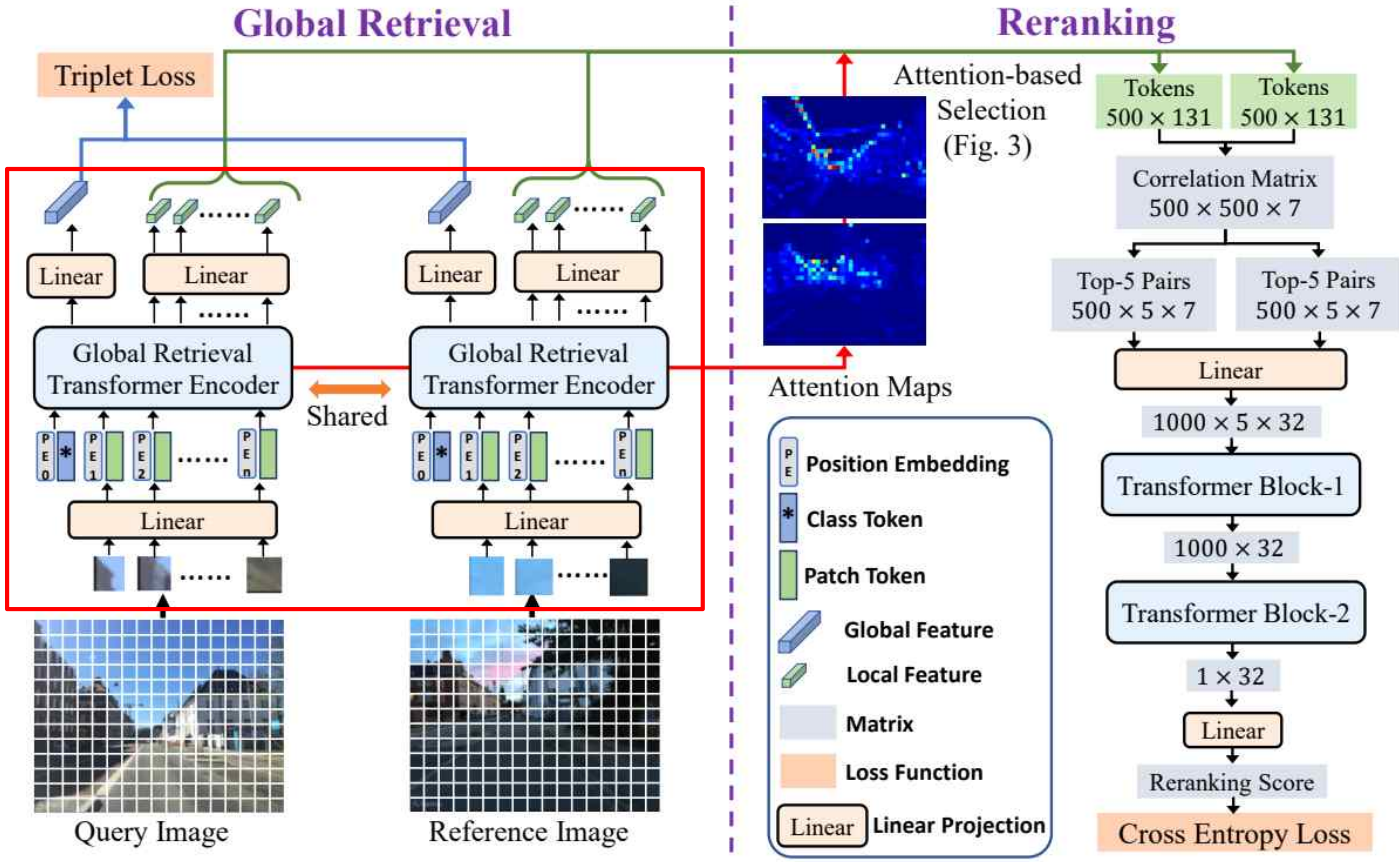
<Qualitative Results>

부산대학교 PUSAN NATIONAL UNIVERSITY
CVSP Lab Computer Vision & Signal Processing

R2Former: Unified Retrieval and Reranking Transformer for Place Recognition[CVPR 2023]

Method

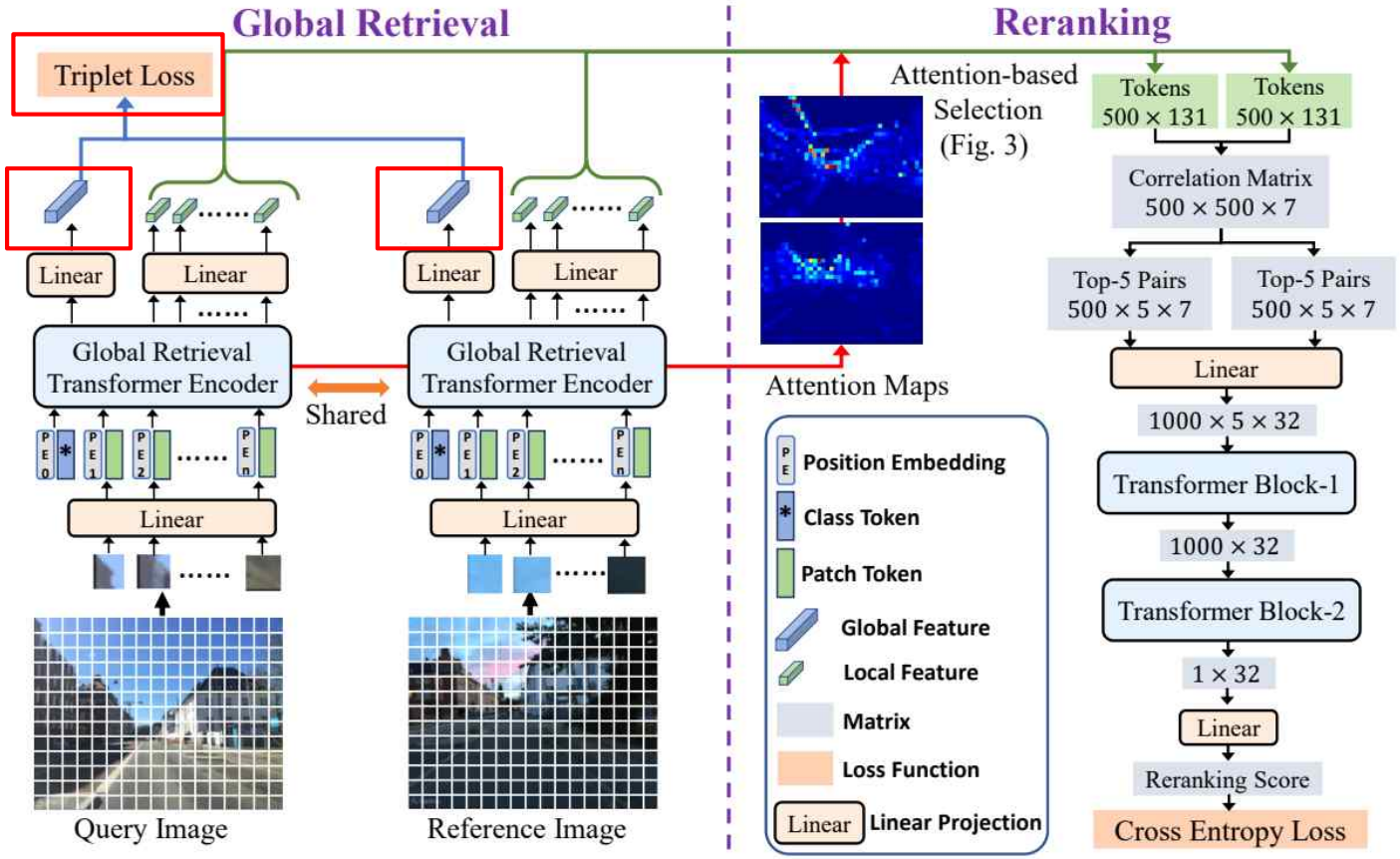
- Feature extraction
  - Global, Local Descriptor 추출
    - Transformer encoder를 활용하여 쿼리 이미지와 데이터베이스 이미지에서 Global·Local Descriptor 추출



<Framework>

Method

- Global Retrieval
  - Top-k 이미지 선별
    - Transformer encoder를 활용하여 얻은 Global Descriptor간의 유사도를 통해 Top-K 이미지 선별



<Framework>

Method

- Global Retrieval
  - Top-k 이미지 선별
    - Triplet Loss를 이용하여 쿼리-양성은 가까이, 쿼리-음성은 멀게 학습 → Top-k 검색 정확도 향상

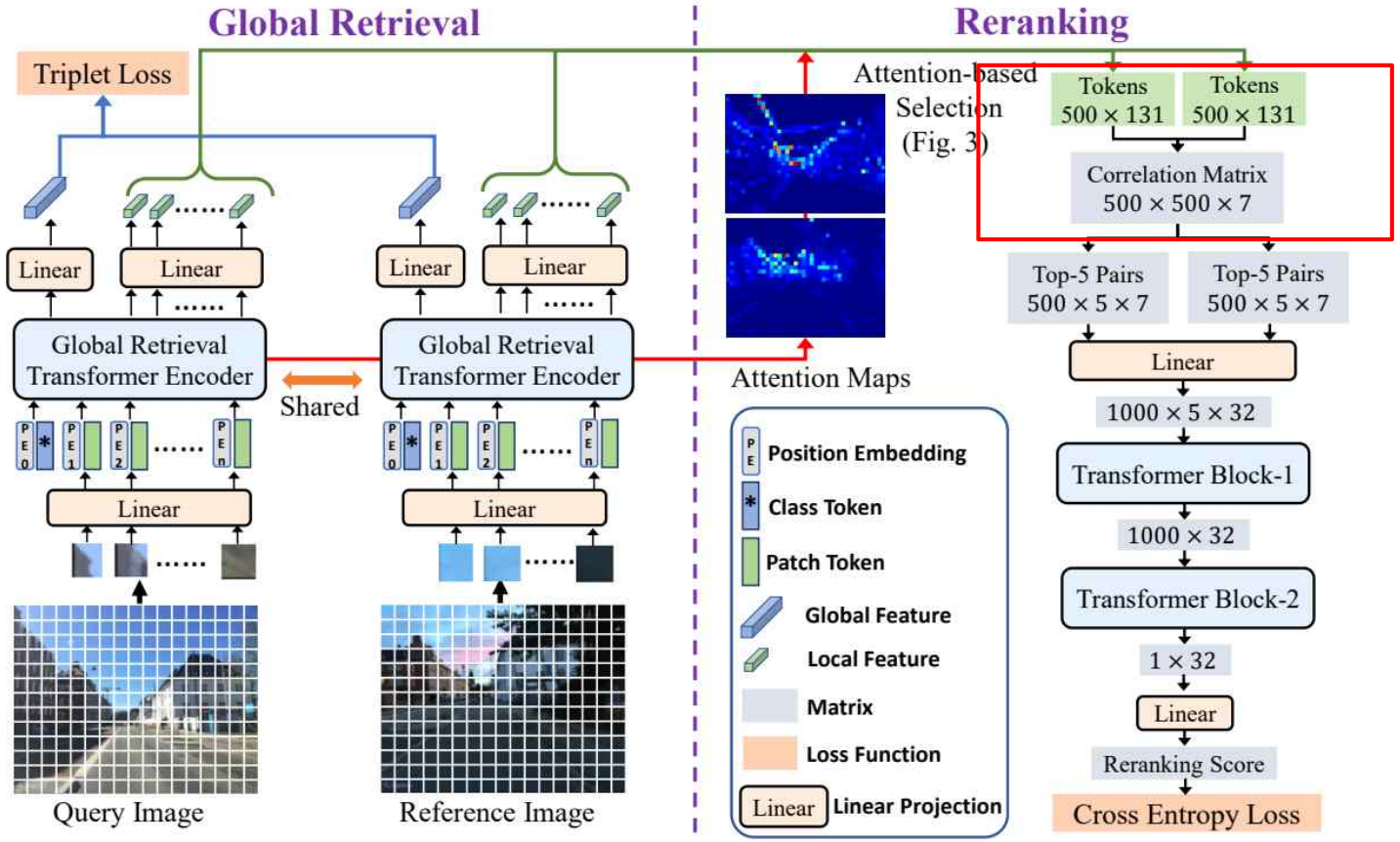
$$L_{retrieval} = max(||E_q - E_p||^2 - ||E_q - E_n||^2 + m, 0)$$

- E_q: 쿼리 이미지의 전역 벡터
- E_p: $Positive$ 이미지의 전역 벡터(같은 장소)
- E_n: $Negative$ 이미지의 전역 벡터(다른 장소)
- m: 마진 값

<Loss 수식>

Method

- Reranking
  - 중요도를 기준으로 500개의 Local Descriptor 선별
    - CLS 토큰과 패치 토큰 간의 Attention 값을 통해 중요하다고 판단되는 500개의 Local Descriptor만 남기기

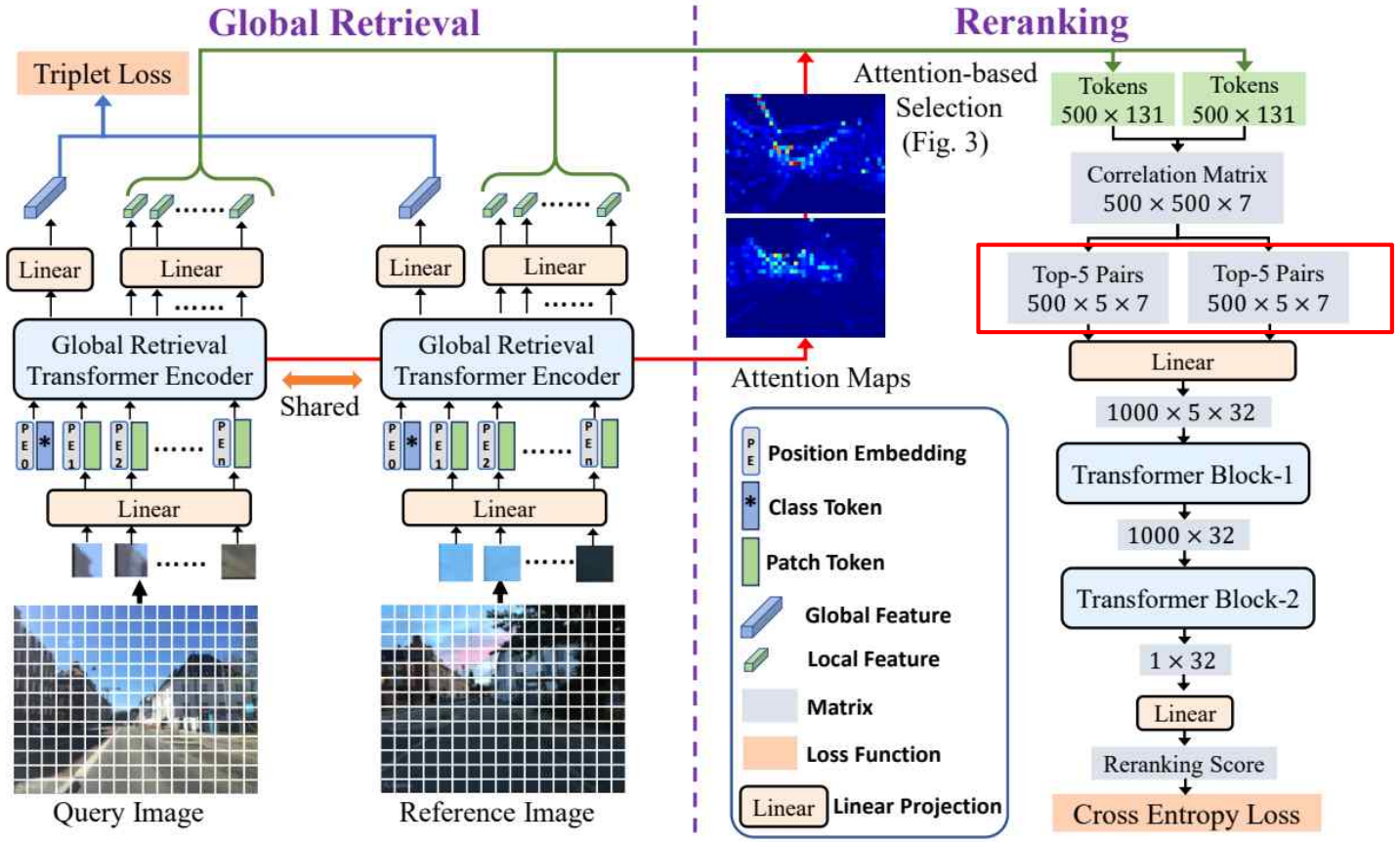


<Framework>

부산대학교 PUSAN NATIONAL UNIVERSITY
CVSP Lab Computer Vision & Signal Processing

Method

- Reranking
  - 유사도 기반 Top-5 매칭
    - 쿼리 이미지와 각 Top-k 후보 이미지에서 추출된 상위 500개의 Local Descriptor를 서로 비교하여 유사도가 높은 5개의 database Local Descriptor만 남김 반대도 적용

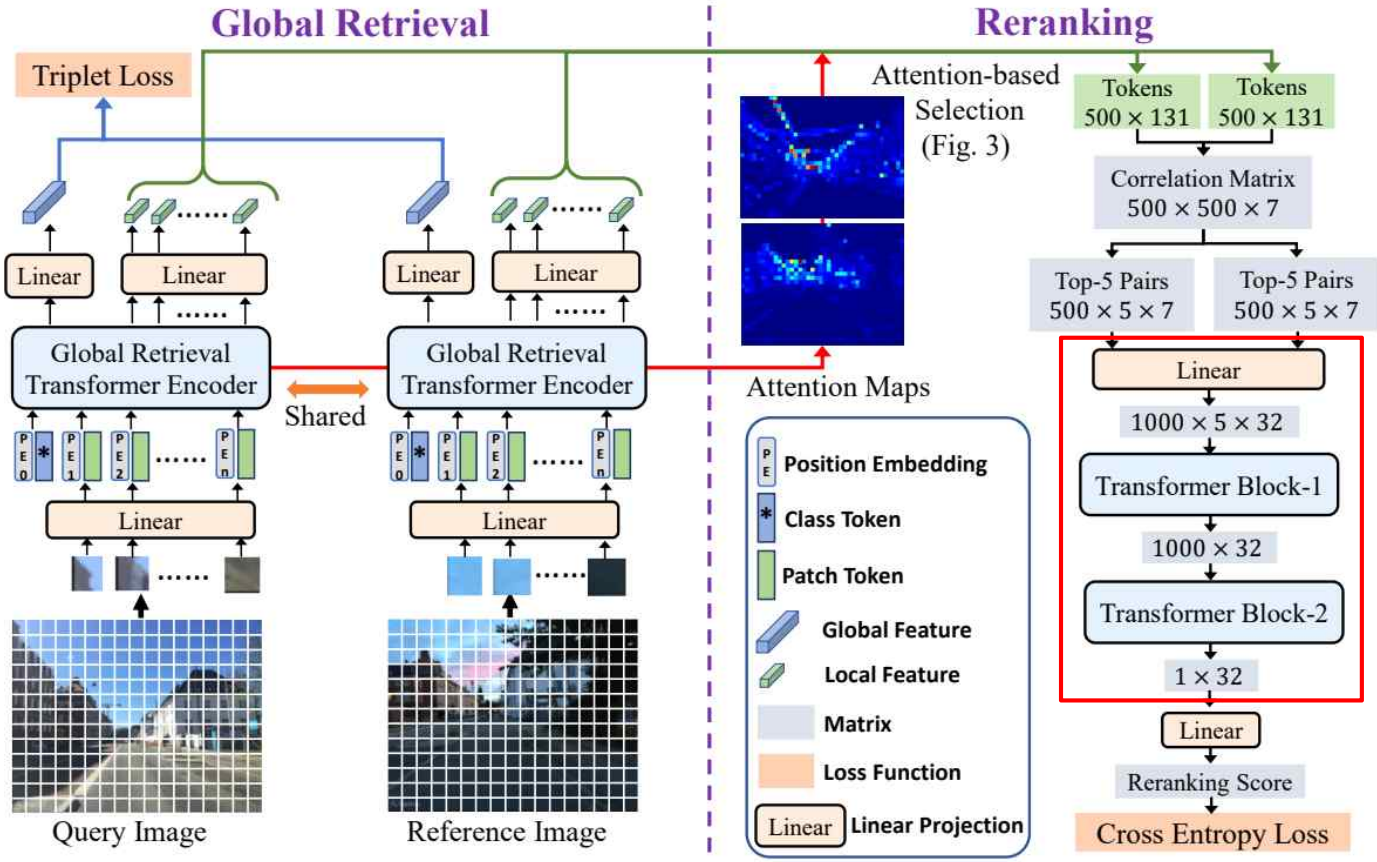


<Framework>

부산대학교 PUSAN NATIONAL UNIVERSITY
CVSP Lab Computer Vision & Signal Processing

Method

- Reranking
  - Use Transformer Block
    - 남은 Local Descriptor를 Transformer Block에 집어넣어 처리하여 후보 이미지와 쿼리 이미지 간의 정교한 매칭 정보가 인코딩된 벡터를 생성(self-attention이후 cross-attention)

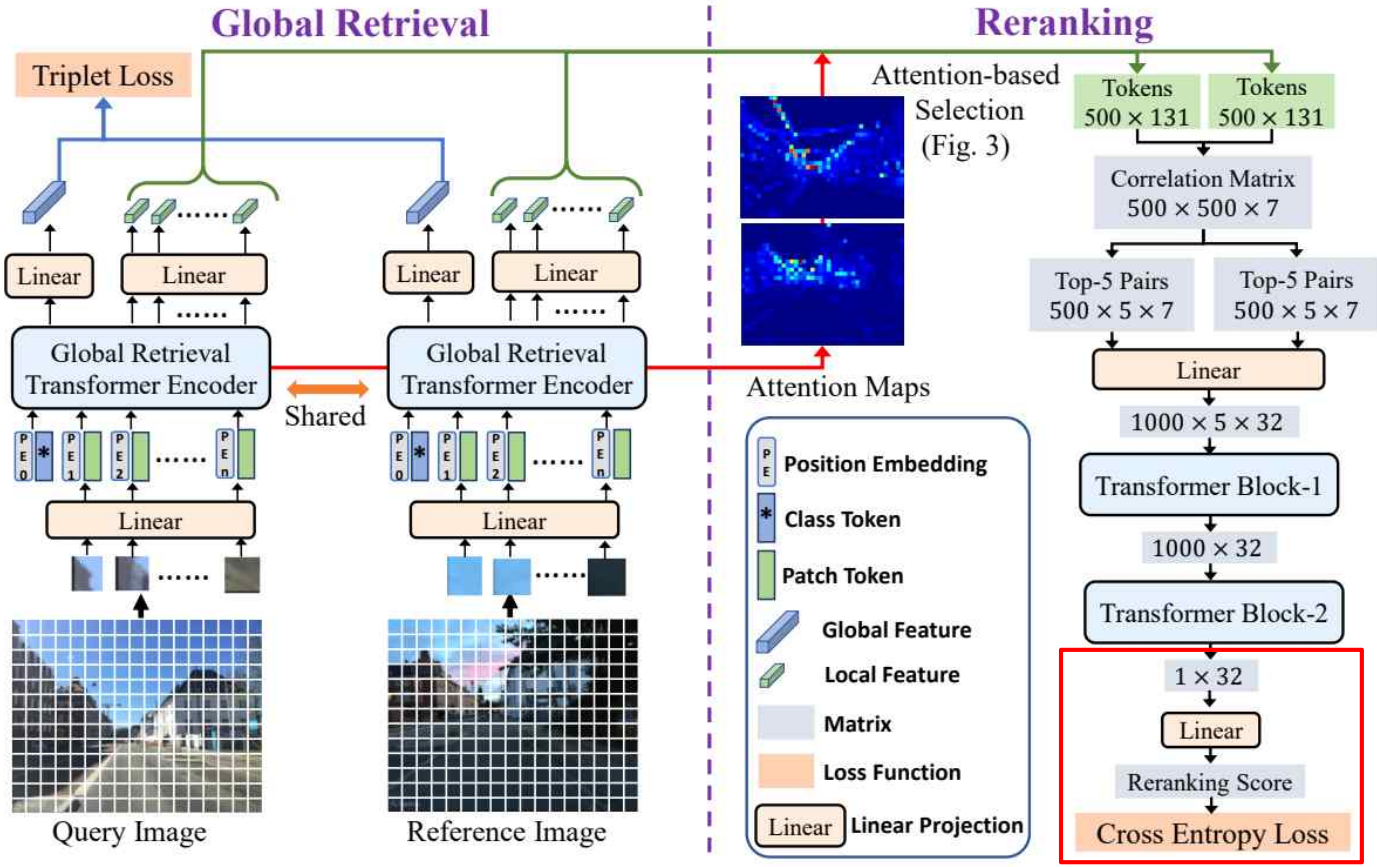


<Framework>

부산대학교 PUSAN NATIONAL UNIVERSITY
CVSP Lab Computer Vision & Signal Processing

Method

- Reranking
  - Calculate similarity and Reranking
    - Transformer 출력 벡터를 평균 풀링하고 FC Layer를 거쳐 유사도 점수를 계산 이후 Top-K 이미지를 이 점수를 기준으로 Reranking



<Framework>

Method

- Training and Loss function
  - Transformer 기반의 Reranking Module을 학습
    - Reranking Module(Transformer+FC)로구한 Similarity Score[0,1]와 정답(1)과 오답(0)을 비교하여 Binary Cross-Entropy (BCE) Loss를 통해 Reranking Module내의 파라미터를 학습
    - 이때, Global retrieval 부분의 Transformer는 Frozen 시킨채로 학습시킴

$$L = -\frac{1}{N}\sum_{i=1}^{N}[y_i \log(\sigma(z_i)) + (1-y_i)\log(1-\sigma(z_i))]$$

- z_i: 쿼리 – 후보 이미지 쌍의 유사도 점수
- $y_i \in \{0,1\}$:레이블
  - $y = 1$:쿼리와 후보 이미지가 동일한 장소
  - $y = 0$:서로 다른 장소
- $\sigma(z) = \frac{1}{1+e^{-z}}$

<Loss 수식>

Method

- Training and Loss function
  - 전체 학습 파이프라인
    - 1. Triplet Loss를 이용해 Global retrieval 부분의 Transformer를 학습
    - 2. 1번에서 학습시킨 Global retrieval 부분의 Transformer를 얼리고 Binary Cross-Entropy (BCE) Loss를 통해 Reranking Module내의 파라미터를 학습
    - 3. 어려운 예시(partial negative mining)를 통해 두 모듈다 파인튜닝
      - partial negative mining: 정답 쿼리와 가장 가까운 100개의 음성 샘플을 저장해뒀다가 파인튜닝에 사용

Experiment

- Quantiative Results

	MSLS Val [55]			MSLS Challenge [55]			Pitts30k [50]			Tokyo 24/7 [49]		
	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10
NetVLAD [3]	60.8	74.3	79.5	35.1	47.4	51.7	81.9	91.2	93.7	64.8	78.4	81.6
SFRS [23]	69.2	80.3	83.1	41.5	52.0	56.3	89.4	94.7	95.9	85.4	91.1	**93.3**
SP-SuperGlue [15, 44]	78.1	81.9	84.3	50.6	56.9	58.3	87.2	94.8	**96.4**	88.2	90.2	90.2
Patch-NetVLAD [26]	79.5	86.2	87.7	48.1	57.6	60.5	88.7	94.5	95.9	86.0	88.6	90.5
TransVPR [53]	86.8	91.2	92.4	63.9	74.0	77.5	89.0	94.9	96.2	79.0	82.2	85.1
Ours	**89.7**	**95.0**	**96.2**	**73.0**	**85.9**	**88.8**	**91.1**	**95.2**	96.3	**88.6**	**91.4**	91.7

<Quantiative Results>

Experiment

- Quantiative Results

	Feature Dim ↓		Latency per Query (ms) ↓		
	Global	Local	Extraction	Retrieval	Reranking
ResNet101 + NetVLAD [3, 6]	65536	N/A	9.60	2.33	N/A
Patch-NetVLAD-s [26]	512	936 × 512	9.29	0.08	952.85
Patch-NetVLAD-p [26]	4096	2826 × 4096	9.36	0.19	8377.17
TransVPR [53]	**256**	1200 × 256	**6.20**	**0.07**	1757.70
Ours	**256**	**500 × (128+3)**	8.81	**0.07**	**202.37**

<Quantiative Results>

Conclusion

- ETR은 Transformer구조를 이용하여 Global descriptor와 Local descriptor를 동시에 사용하여 빠르고 정확하게 Reranking하는데 성공함

- R2Former는 Global retrieval과 Local reranking을 하나의 Transformer 기반 구조로 통합하여, 장소 인식에서 정확도와 효율성을 동시에 달성함

- 앞으로 이 두 논문의 인용작(후속작)들을 Survey해가며 과제 수행에 필요한 지식을 쌓아갈 예정임

감사합니다