Improving Spatial Context of Global Descriptors for Image Retrieval

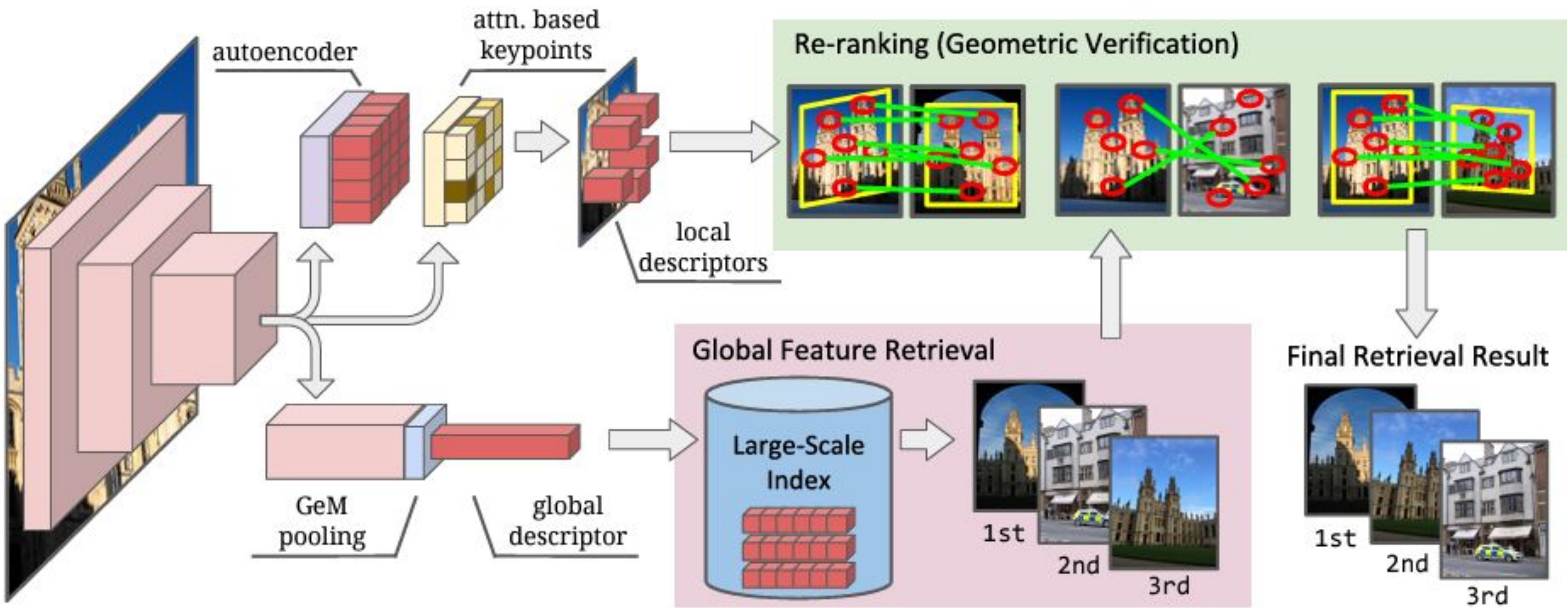
Jinhwan Seo, Kyu Beom Han
(서진환, 한규범)

2024/05/01

KAIST

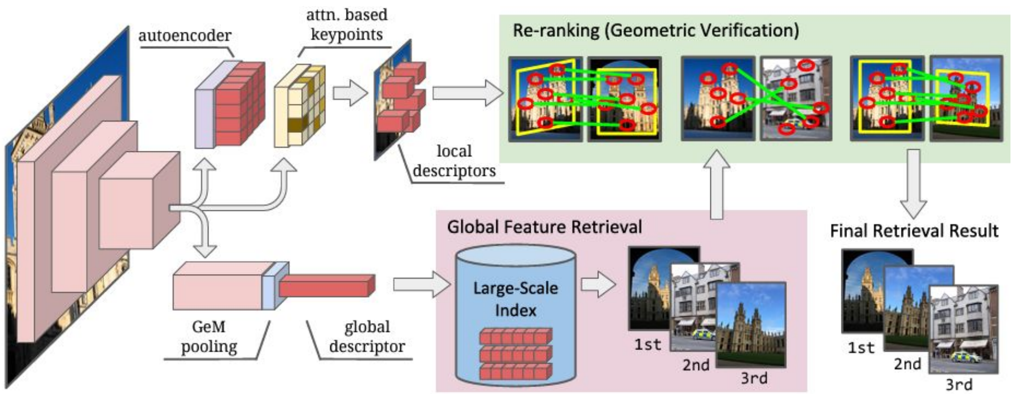
Overall Retrieval Framework

- Global descriptors for efficient ranking
- Local descriptors for precise re-ranking based on geometric similarity



Motivation

- Re-ranking (e.g., spatial verification) is necessary because ranking via global descriptors often lack spatial context between local features (descriptors)
- Increasing initial search performance can reduce necessity of re-ranking, making retrieval efficient
- Add spatial context to global descriptors



(b) Breakdown of average time per query.

very slow

initial search	hypergraph propagation	uncertainty calculation	spatial verification
0.62 s	1.07 s	0.0003 s	$\boxed{41.12 \text{ s}}$

Slide from CS588

Brief Idea

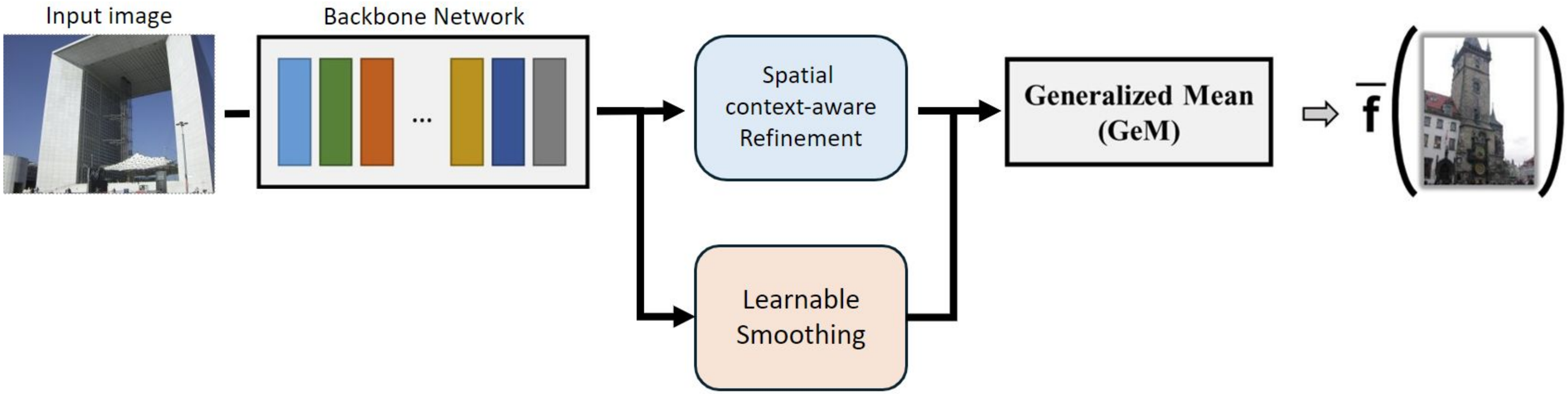
Provide <u>spatial context between local descriptors</u> for global descriptor

Method 1: Learnable Smoothing

- Local spatial context

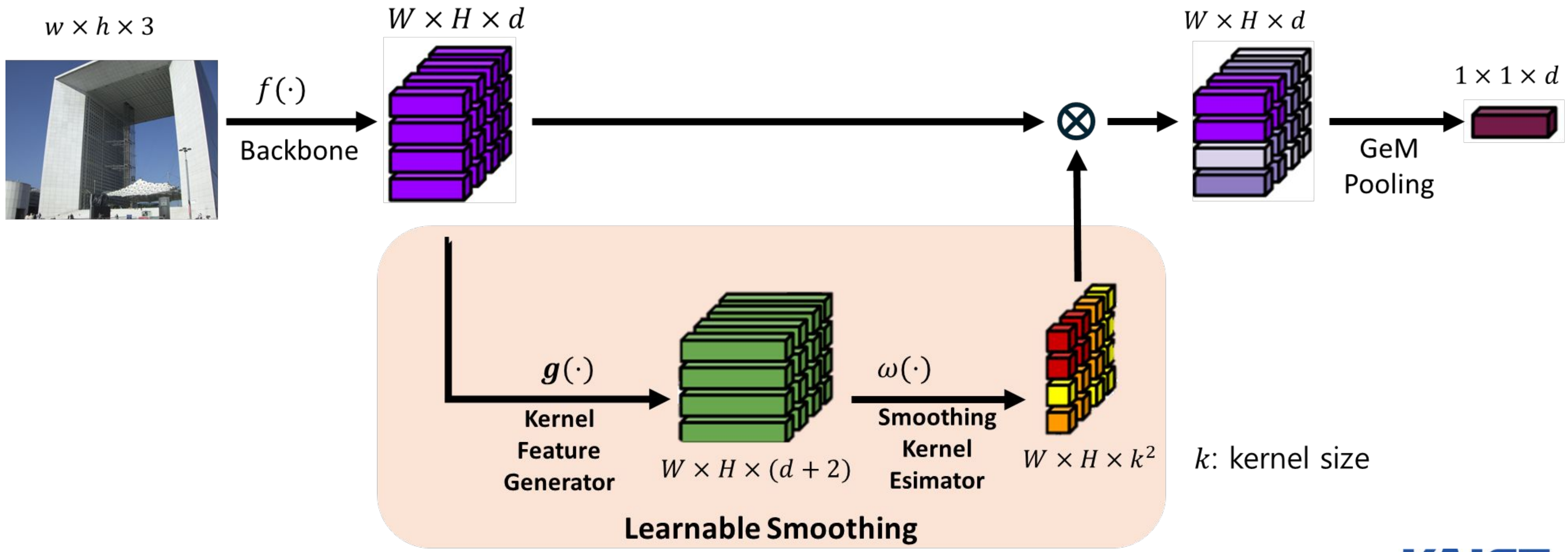
Method 2: Spatial context-aware refinement

- Global spatial context



Method 1. Learnable Smoothing

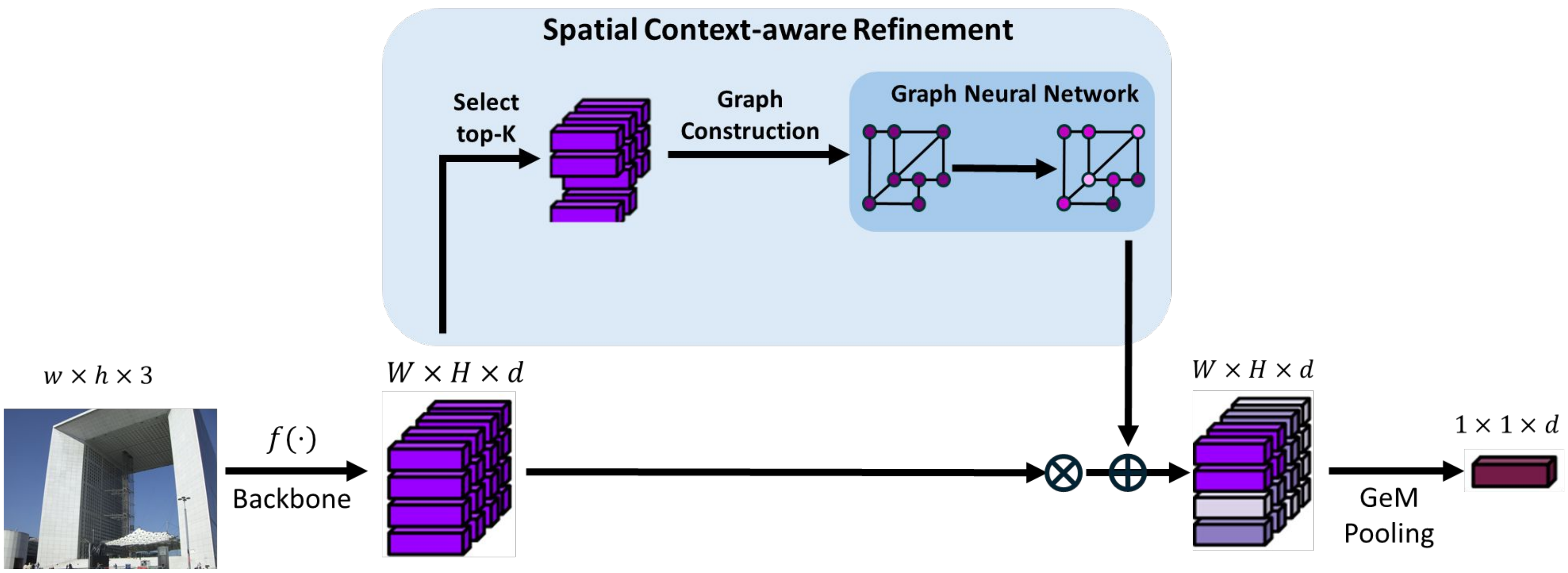
- Gather neighbor local feature based on learned weight
- <u>Learnable kernel bandwidth</u> (receptive field) for smoothing
- Estimate <u>self-confidence</u> to reduce burstiness



k: kernel size

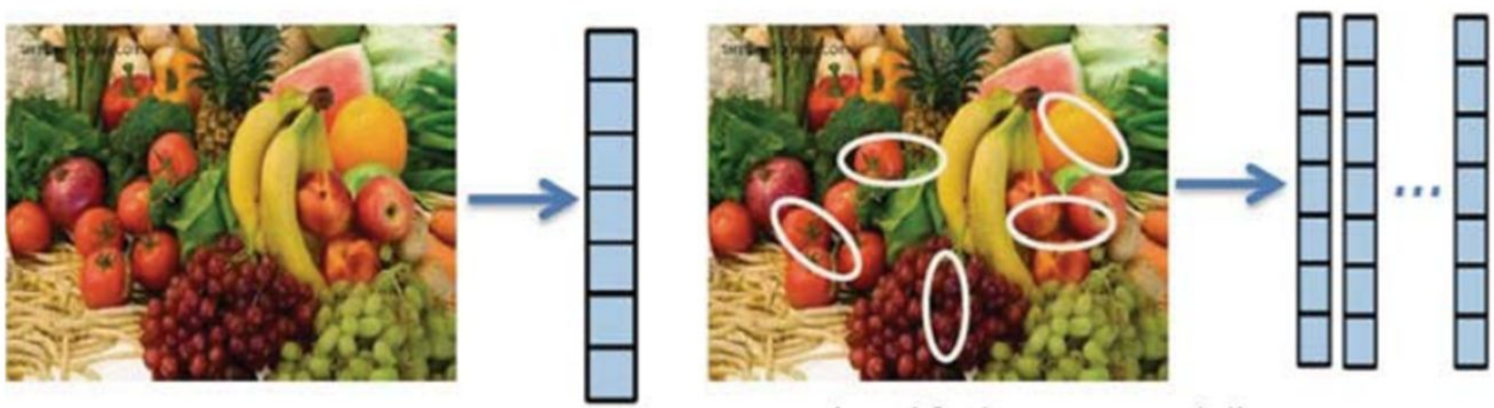
Method 2. Spatial Context-aware Refinement

- Learnable smoothing focus on limited region of locality
- Each GCN propagates messages to next block
- Can consider global spatial and semantic context



Global v.s. Local Descriptors

- Local descriptors
  - Represents multiple keypoints
  - Contains spatial & geometric relationship
  - Exhaustive to match local descriptors between multiple images
- Global descriptors
  - Represents an image
  - Mostly an aggregation of local descriptors
  - Efficient matching between multiple images

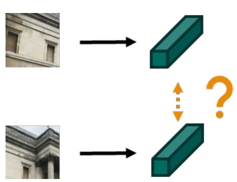


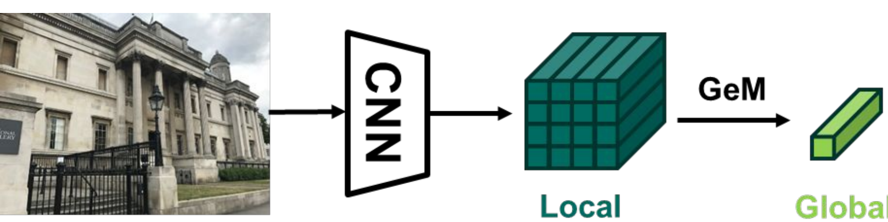
Global feature representation

Local feature representation

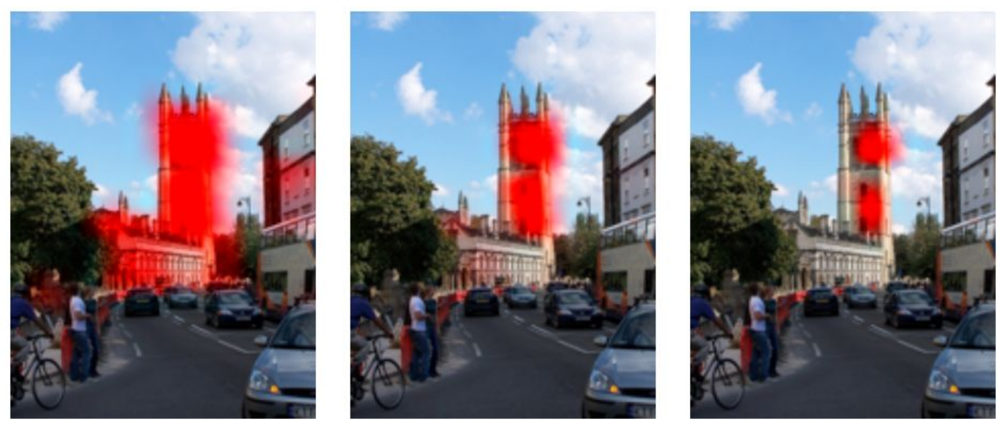
Related Works - GeM

- Generalized Mean Pooling
- Channel-wise Learnable P
- **Limitation**
  - Less control on spatial information





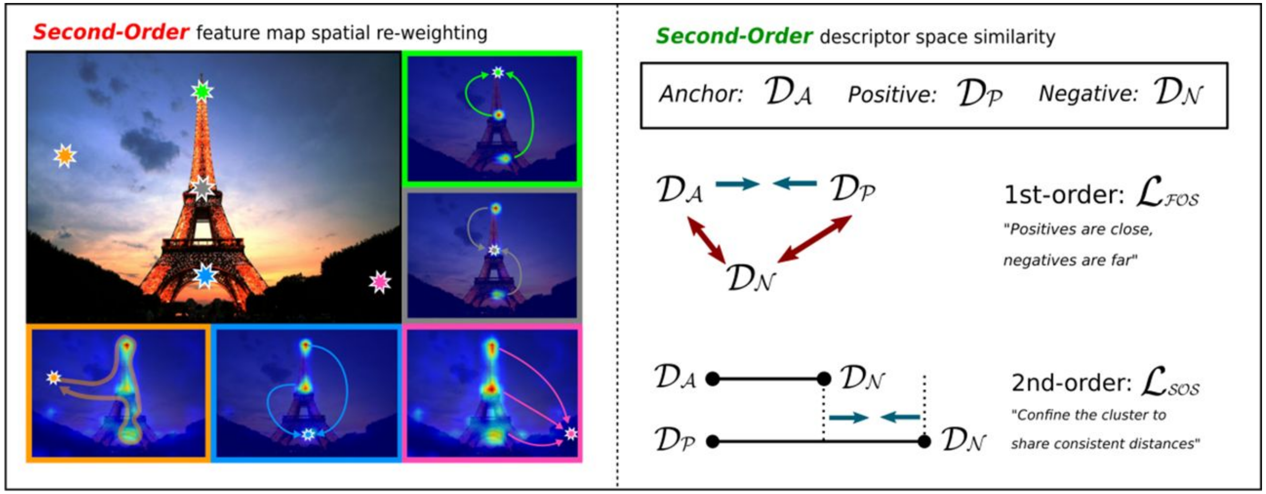
$$\mathbf{f}^{(g)} = [\mathrm{f}_1^{(g)} \dots \mathrm{f}_k^{(g)} \dots \mathrm{f}_K^{(g)}]^\top, \quad \mathrm{f}_k^{(g)} = \left(\frac{1}{|\mathcal{X}_k|}\sum_{x\in\mathcal{X}_k} x^{p_k}\right)^{\frac{1}{p_k}}.$$



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

Related Works - SOLAR

- Re-weighting local descriptor before GeM
- Confine clusters with second-order loss
- **Limitation:**
  - Attention map requires expensive computational cost
  - Cannot guarantee that it contains spatially contextual information

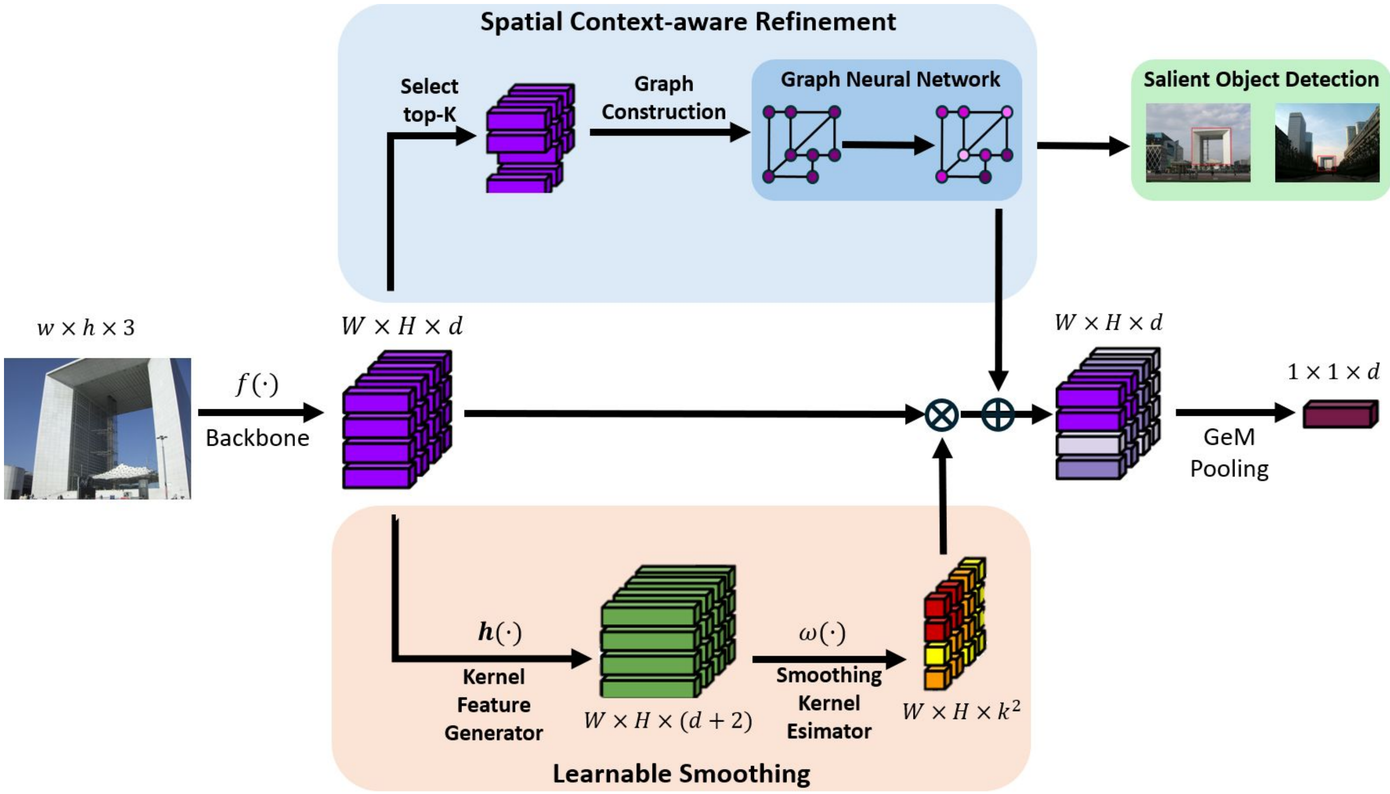


Re-weighting <u>local descriptors</u> prior to GeM

Distribute <u>global descriptors</u> in descriptor space

Overall Method

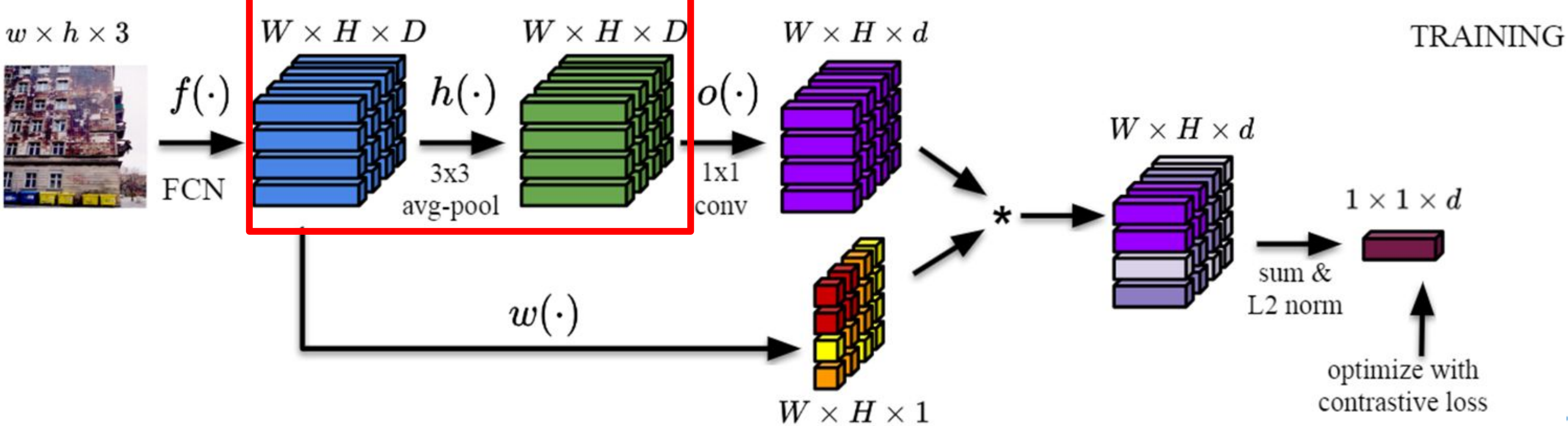
- Learnable smoothing : <u>Local spatial context</u>
- Spatial Context-aware Refinement : <u>Global spatial context</u>



Method 1. Learnable Smoothing

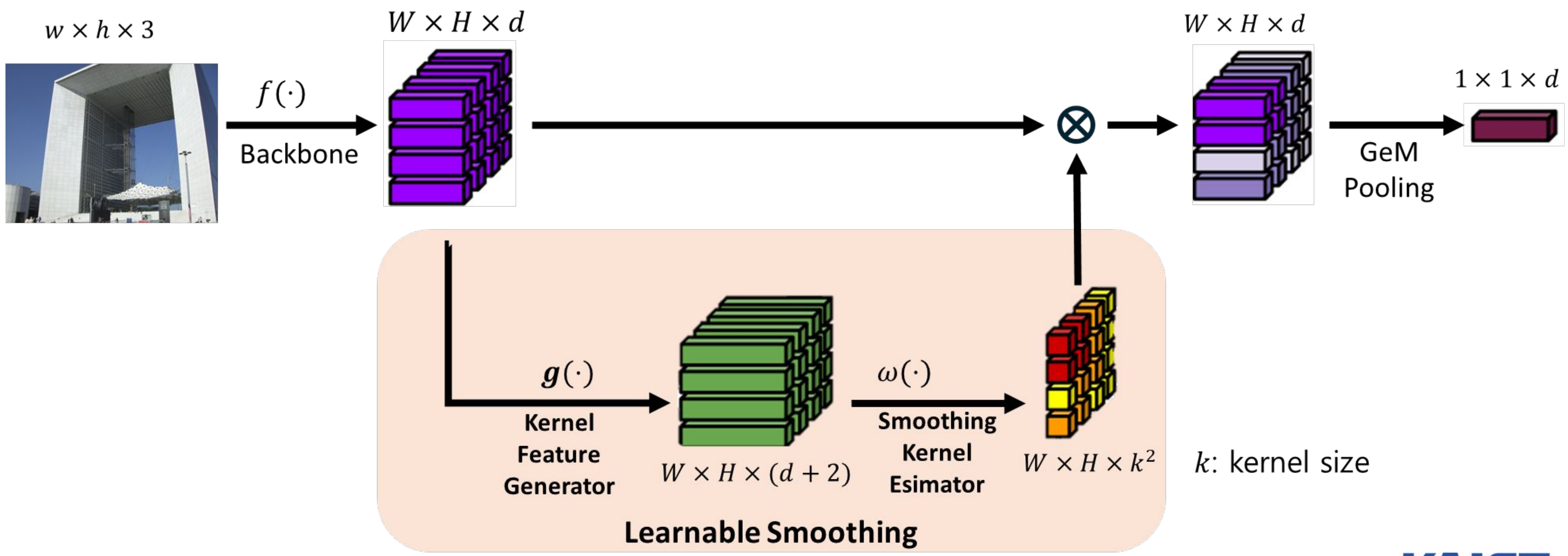
- Smoothing (i.e., AvgPool) <u>adds spatial context & reduces burstiness of local features</u>
- More sophisticated smoothing can provide finer spatial context

Method	Loss	Validation		$\mathcal{R}$Oxford		Tiny-GLD$_2$		
		mAP	$\lvert\mathcal{C}_{\mathcal{X}}\rvert$	mAP	$\lvert\mathcal{C}_{\mathcal{X}}\rvert$	μAP	$\lvert\mathcal{C}_{\mathcal{X}}\rvert$	
5: R18$_{\hat{h}\hat{o}\hat{w}}$	CE	$75.5_{\pm 1.3}$	$391.0_{\pm 8.2}$	$63.7_{\pm 1.6}$	$442.3_{\pm 9.7}$	$64.0_{\pm 1.8}$	$427.5_{\pm 15.6}$	w/o smoothing
8: R18$_{h\hat{o}\hat{w}}$	CE	$77.0_{\pm 0.9}$	$279.6_{\pm 5.6}$	$65.4_{\pm 0.5}$	$320.6_{\pm 6.8}$	$68.6_{\pm 1.8}$	$300.9_{\pm 11.4}$	w/ smoothing



Method 1. Learnable Smoothing

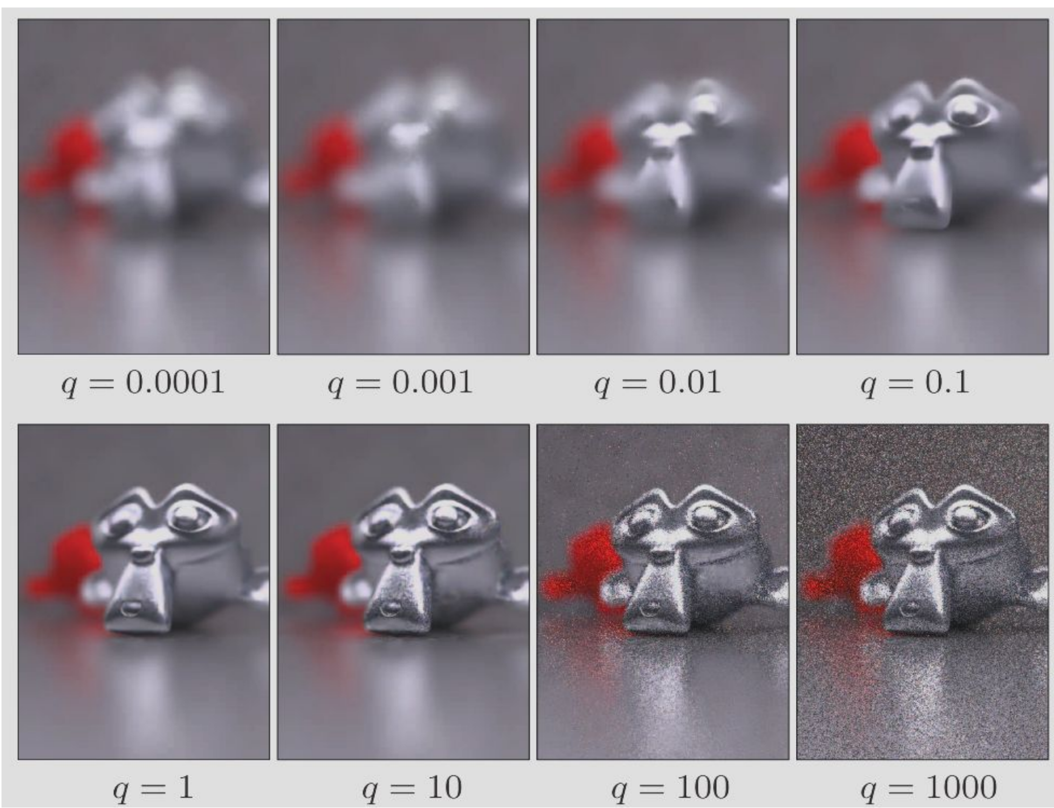
- Gather neighbor local feature based on learned weight
- <u>Learnable kernel bandwidth</u> (receptive field) for smoothing
- Estimate <u>self-confidence</u> to reduce burstiness

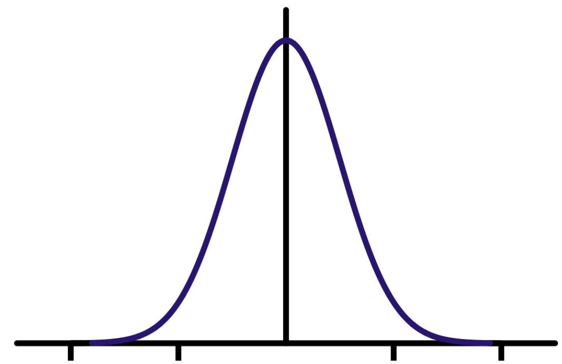


k: kernel size

Prelim. Gaussian Image Filter

- Gaussian image filter reduces high-frequency details (e.g., noise)
- Further developed for image denoising (e.g., bilateral filter)



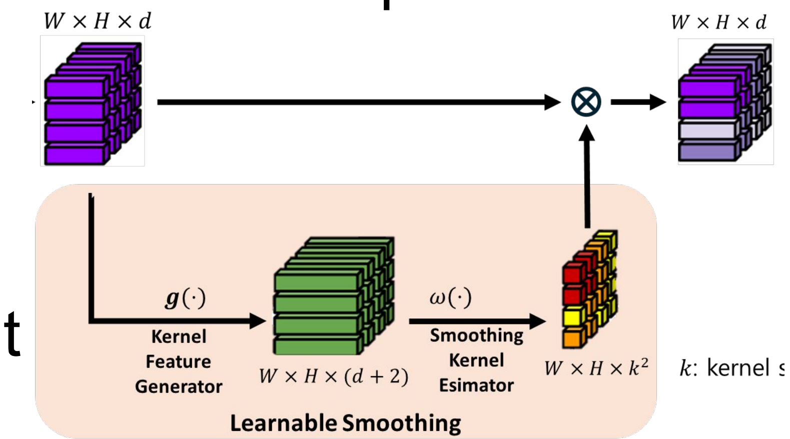


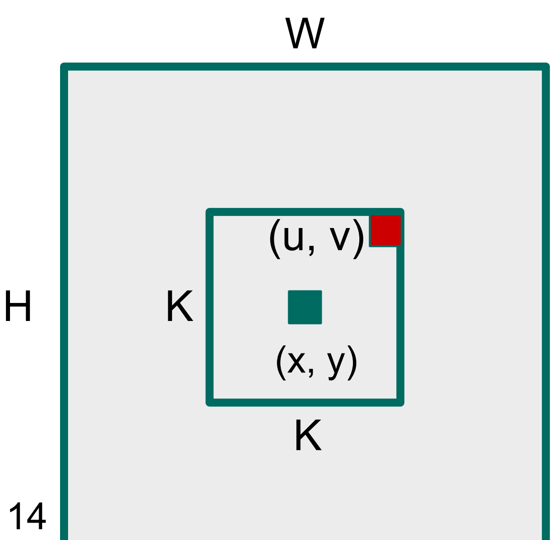
$$g(x, y) = \frac{1}{2\pi q} \exp\left[-q(x^2 + y^2)\right]$$

2d gaussian

Learnable Smoothing Kernel

- Kernel Feature Generator $\boldsymbol{g}(\cdot)$ estimates three pixel-wise features
  - Kernel feature $f_{xy} \in \mathbb{R}^d$
  - Bandwidth $a_{xy} \in \mathbb{R}$
  - Self-confidence $c_{xy} \in \mathbb{R}$
- Calculate Gaussian kernel weight $\omega(\cdot)$ over K x K neighbors

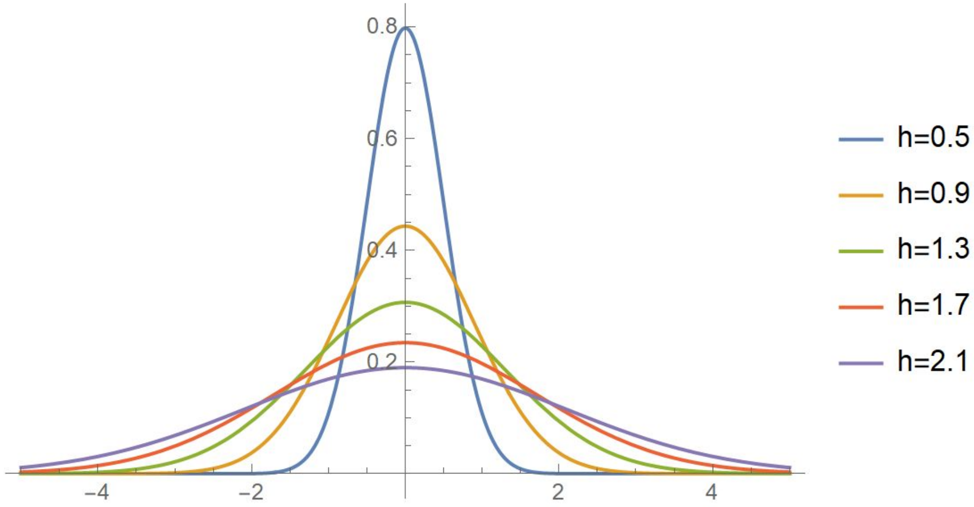




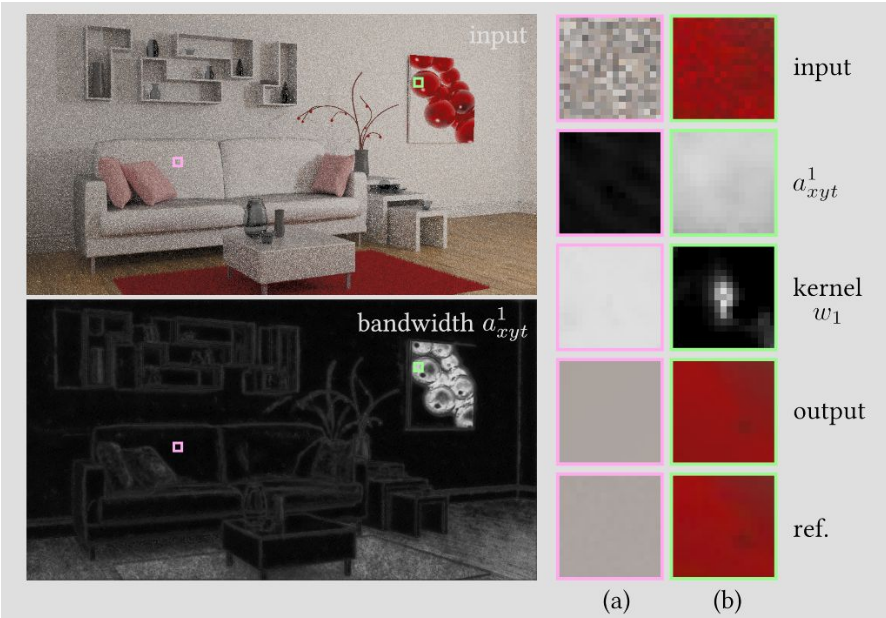
$$w_{xyuv} = \begin{cases} c_{xy} & \text{if } x = u \text{ and } y = v \\ \exp\left(-a_{xy} \left\| f_{xy} - f_{uv} \right\|_2^2\right) & \text{otherwise} \end{cases}$$

Learnable Bandwidth

- Gaussian kernel with learnable bandwidth $-a_{xy}$
  - Larger bandwidth (Narrow): Spatially non-correlated info.
  - Smaller bandwidth (Wide): Spatially correlated info.



Identify high-frequency region for denoising



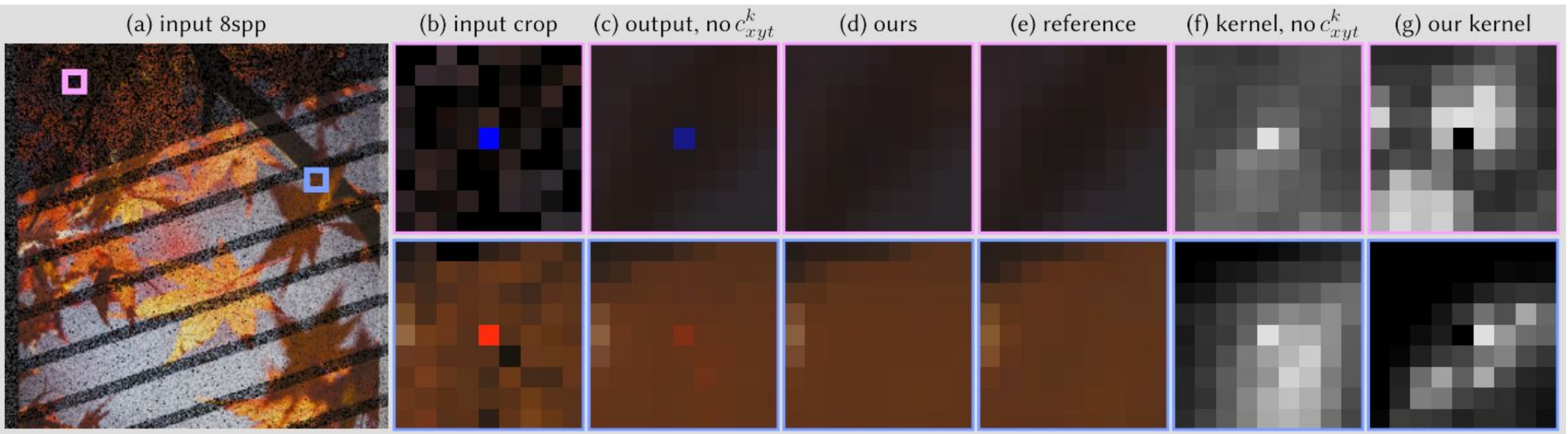
$$w_{xyuv} = \begin{cases} c_{xy} & \text{if } x = u \text{ and } y = v \\ \boxed{\exp\left(-a_{xy} \|f_{xy} - f_{uv}\|_2^2\right) \quad \text{otherwise}} \end{cases}$$

- Kernel feature $f_{xy} \in \mathbb{R}^d$
- Bandwidth $a_{xy} \in \mathbb{R}$
- Self-confidence $c_{xy} \in \mathbb{R}$

Learnable Self-confidence

- Allows to reject itself when it is non-relevant for image retrieval

Helps to reject when the center pixel is an outlier for denoising

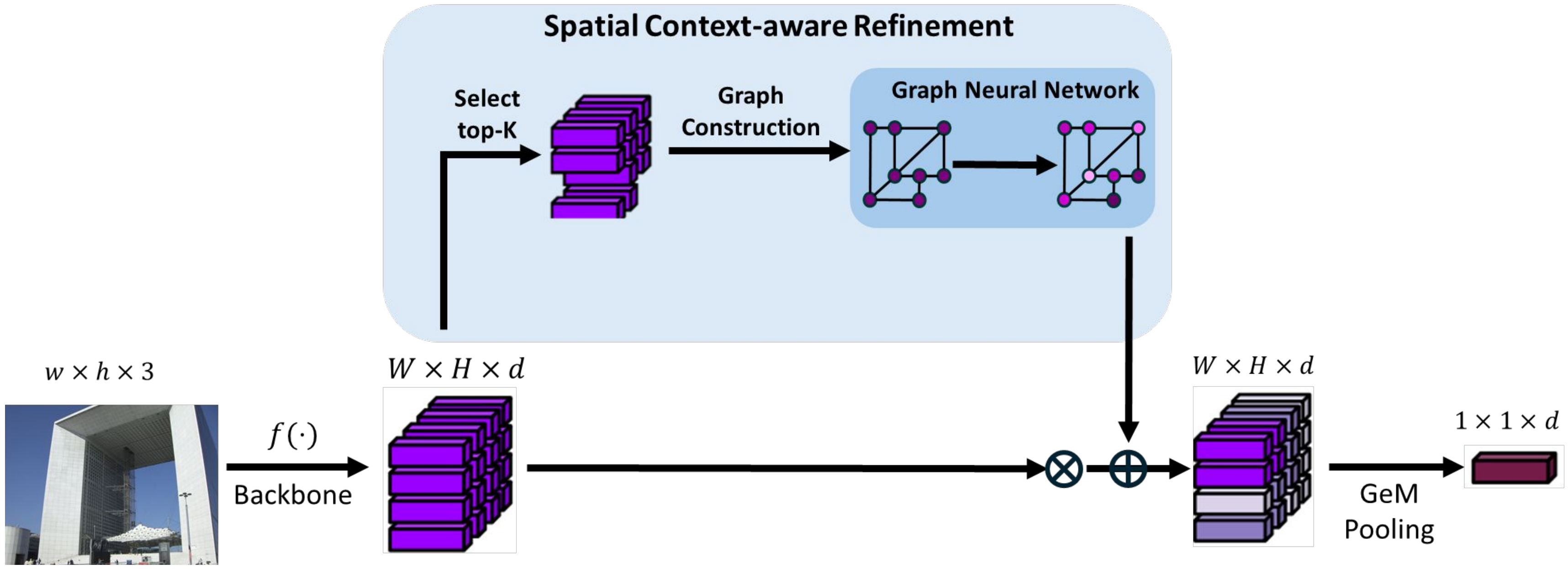


$$w_{xyuv} = \begin{cases} \boxed{c_{xy} \quad \text{if } x = u \text{ and } y = v} \\ \exp\left(-a_{xy} \|f_{xy} - f_{uv}\|_2^2\right) & \text{otherwise} \end{cases}$$

- Kernel feature $f_{xy} \in \mathbb{R}^d$
- Bandwidth $a_{xy} \in \mathbb{R}$
- Self-confidence $c_{xy} \in \mathbb{R}$

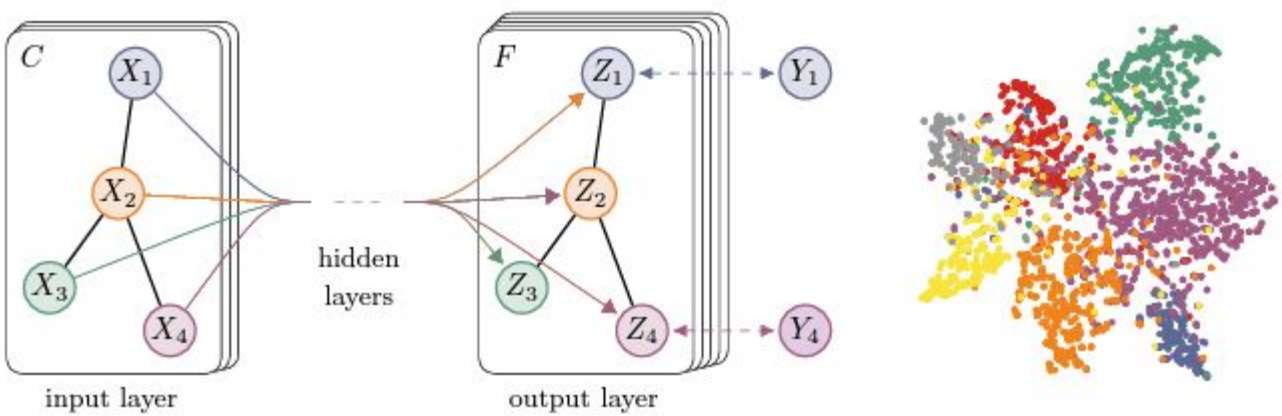
Method 2. Spatial Context-aware Refinement

- Learnable smoothing focus on limited region of locality
- Each GCN propagates messages to next block
- Can consider global spatial and semantic context



Related Works - GNN

- Propagate messages via GCN
- Learn connection between contiguous nodes
- Image retrieval
    - Each local descriptor can be represented as node
    - The spatial relation between each node can be represented as edge

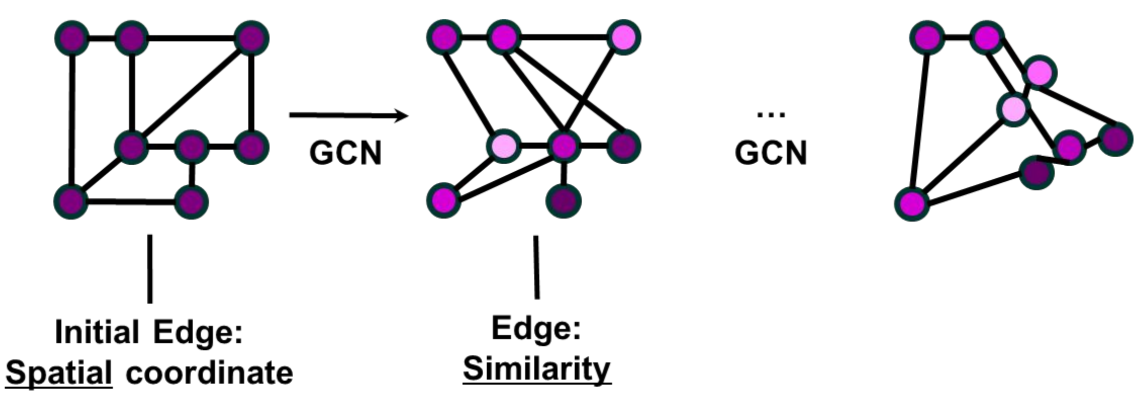


(a) Graph Convolutional Network

(b) Hidden layer activations

GNN for Image Retrieval

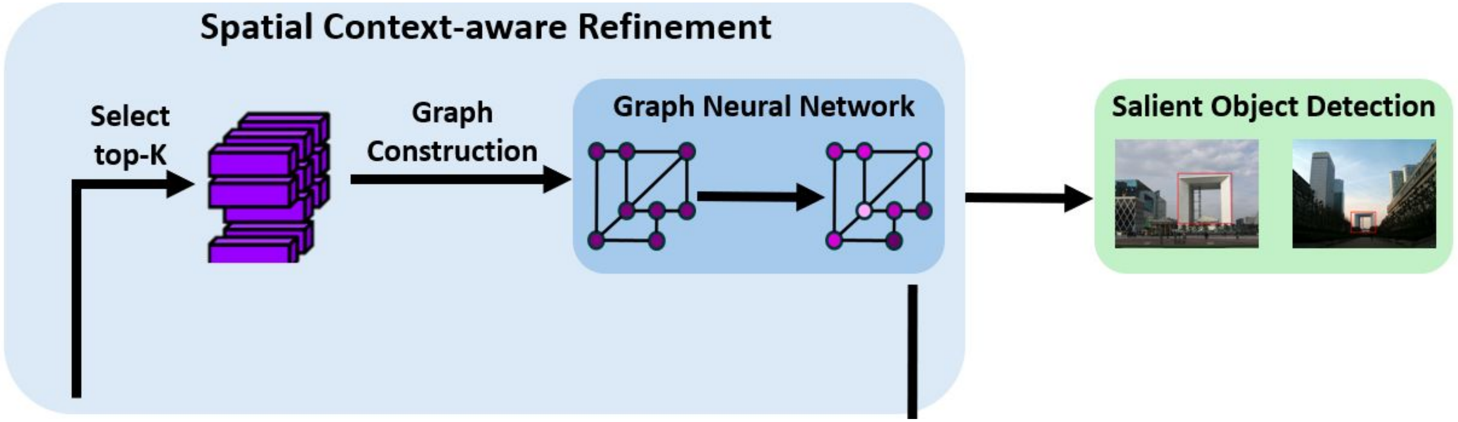
- Create a graph structure for image retrieval
  - Node: Local descriptors (Top-K local features)
  - Edge: Spatial coordinate, similarity
- Extract feature via GNN
  - Context-aware local descriptor
  - Local descriptor could contain not only <u>spatial</u> information but also <u>semantic</u> relations



Verification for localization

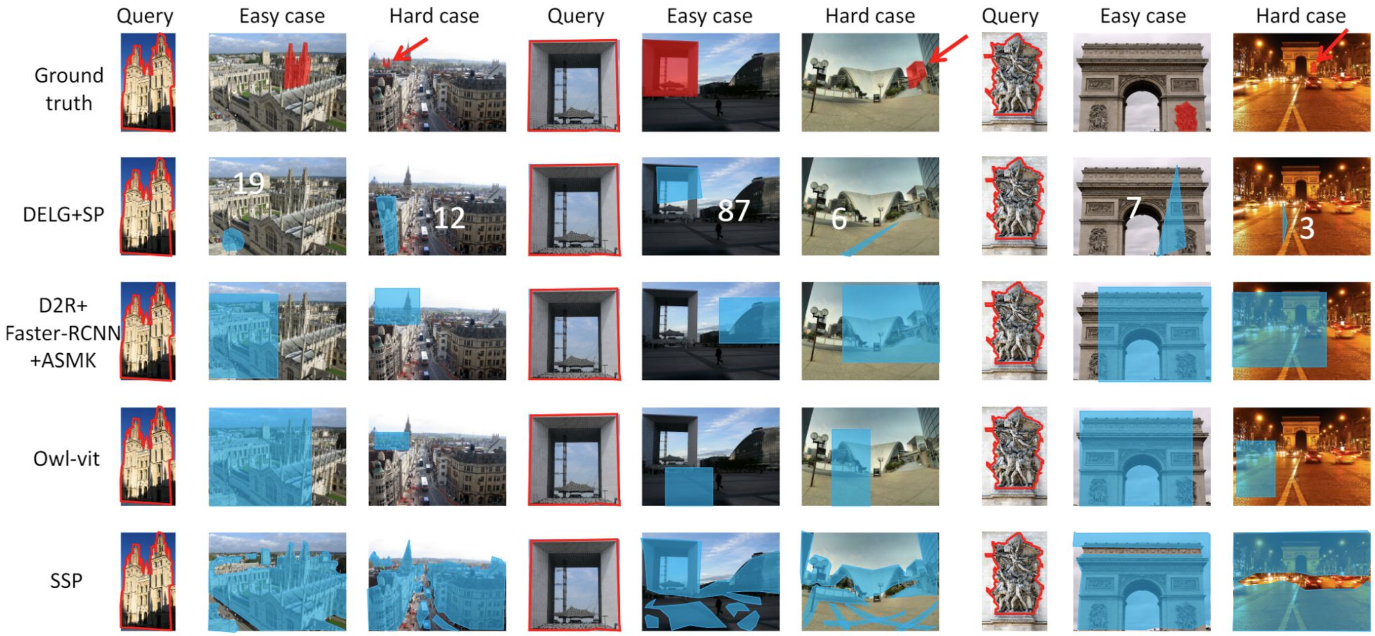
- We have a rich global descriptor using learnable smoothing and spatial context-aware refinement
- How can we evaluate effectiveness of spatial context?

- We extend salient object detection to verify spatial context as localization performance



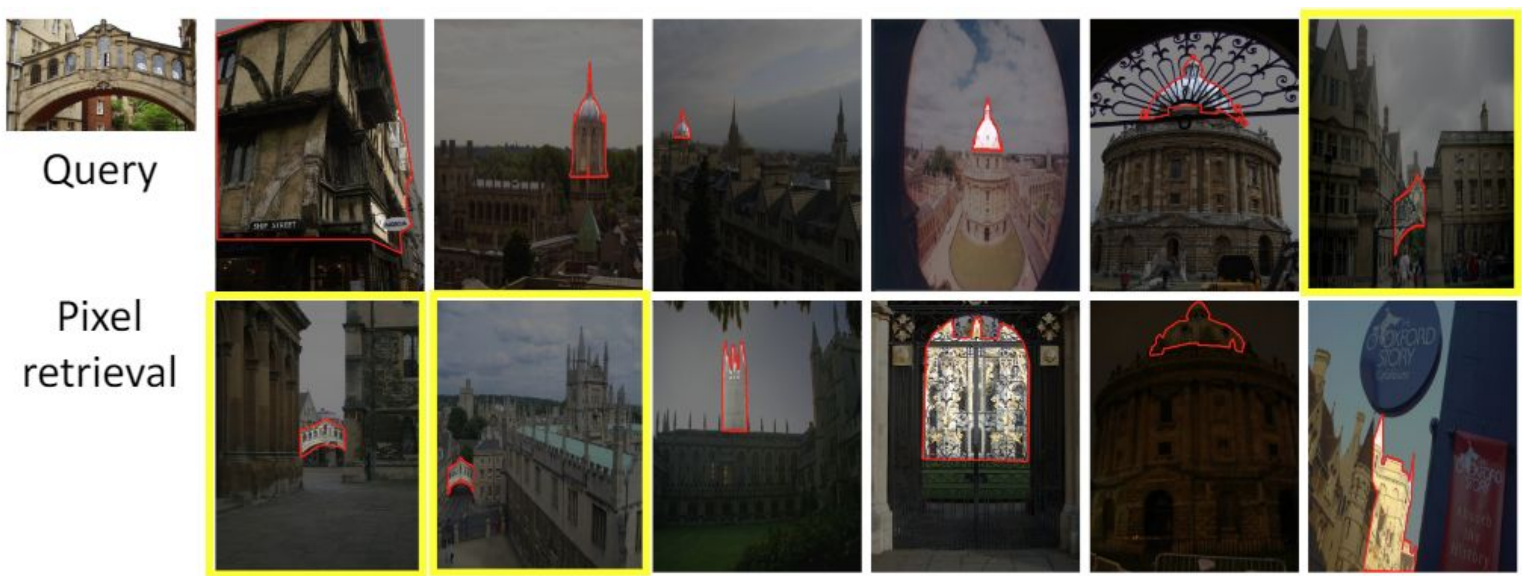
Related Works - Pixel retrieval

- Evaluate localization methods for pixel retrieval
  - Spatial Verification: SIFT, DELF and DELG
  - One-shot Detection: Faster R-CNN, SSD and D2R
  - Dense matching: GLUNet, WarpC, …



Auxiliary. Salient Object Detection

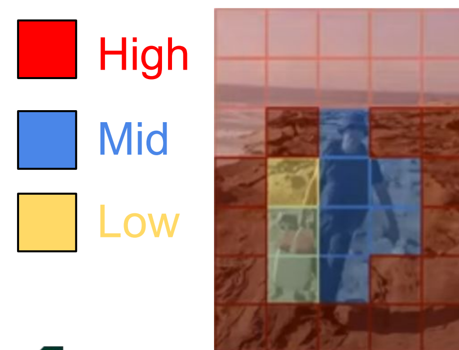
- Pixel retrieval
  - Query-based interaction
  - Enhance user experience in retrieval results
- Salient Object Detection
  - Identify foreground / distinctive part of objects in intra-image

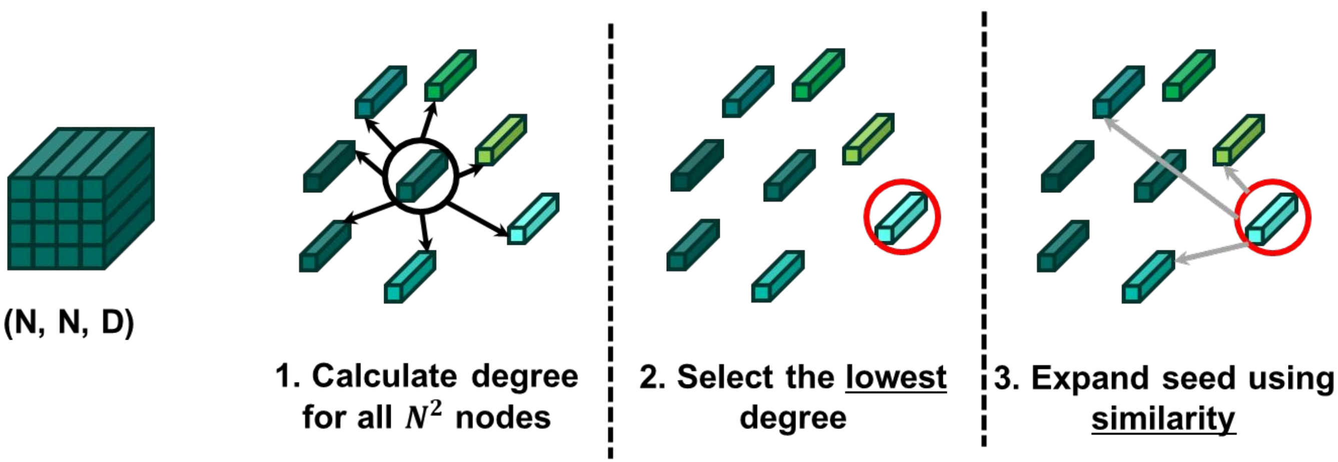


Localizing Objects with Self-Supervised Transformers and no Labels, Simeoni., BMVC 2021

Related Works - SOD

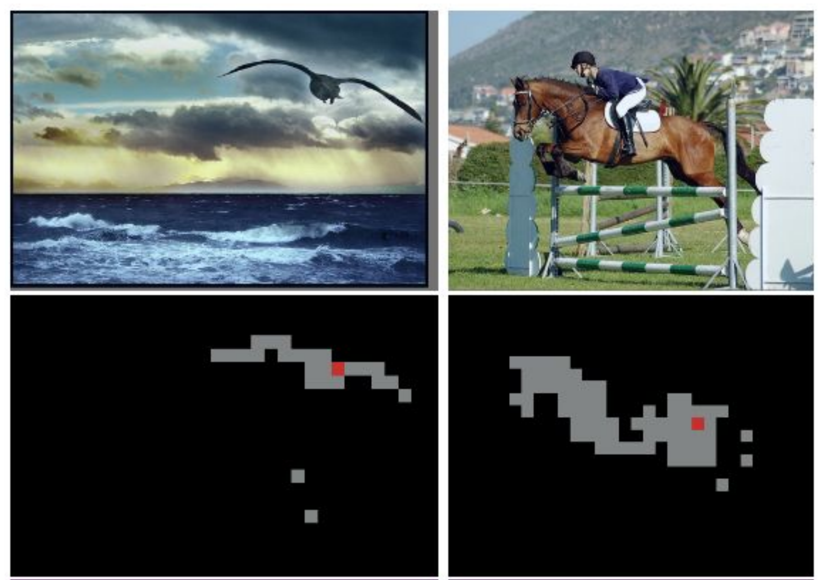
- Unsupervised object discovery
- Assumption:
  - Foreground features are less correlated than background
  - Less features of foreground than background
- Method:
  - Use the information of degree
  - Object seed: patch with the lowest degree
  - Expand features similar with object seed





Related Works - SOD

- Pros
  - Quick (60FPS)
  - Simple and effective
- Cons
  - Single object detection
  - Issues when object covers most of image

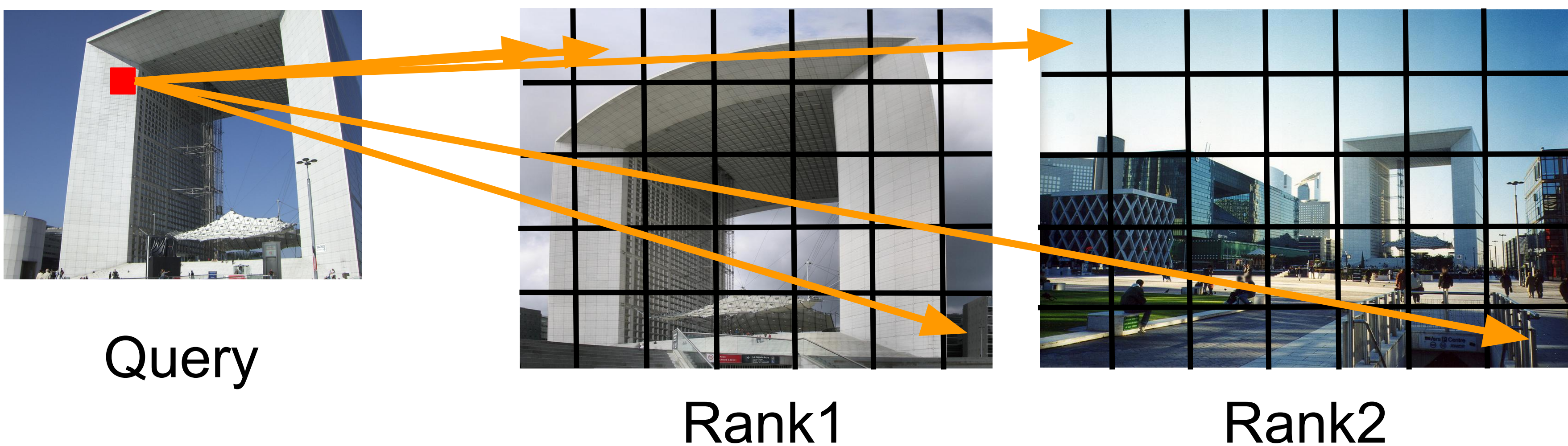


Object seed

Expand patches

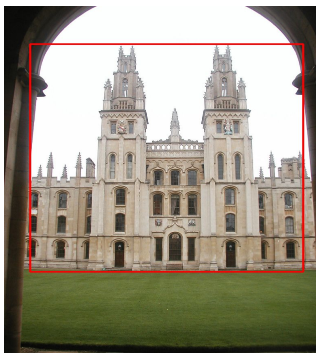
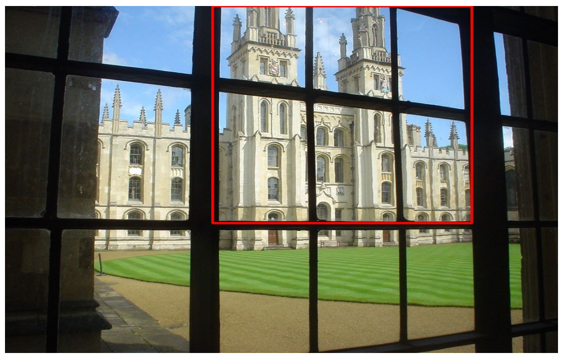
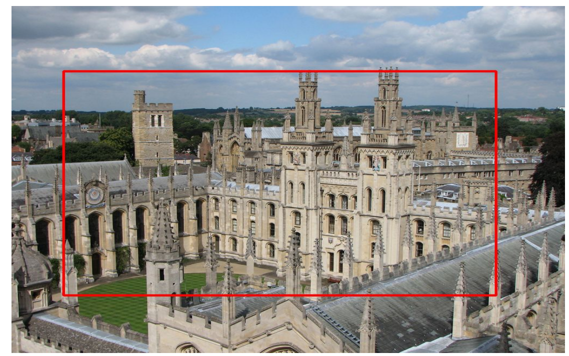
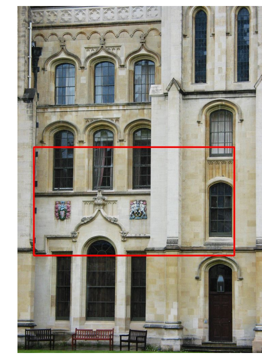
Auxiliary. Salient Object Detection

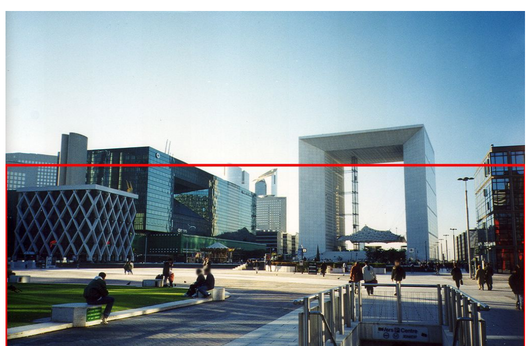
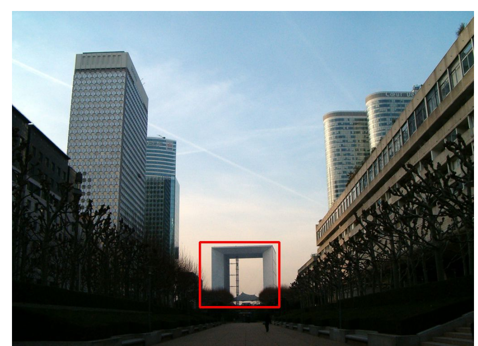
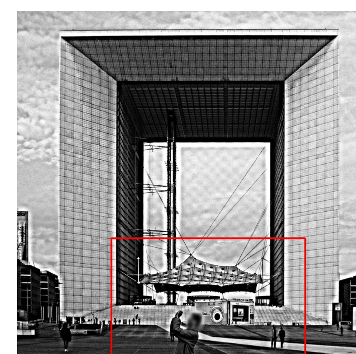
- Query-based interaction
  a. Select initial seed in query image
  b. Calculate similarity in each gallery image (Expansion)



Auxiliary. Salient Object Detection

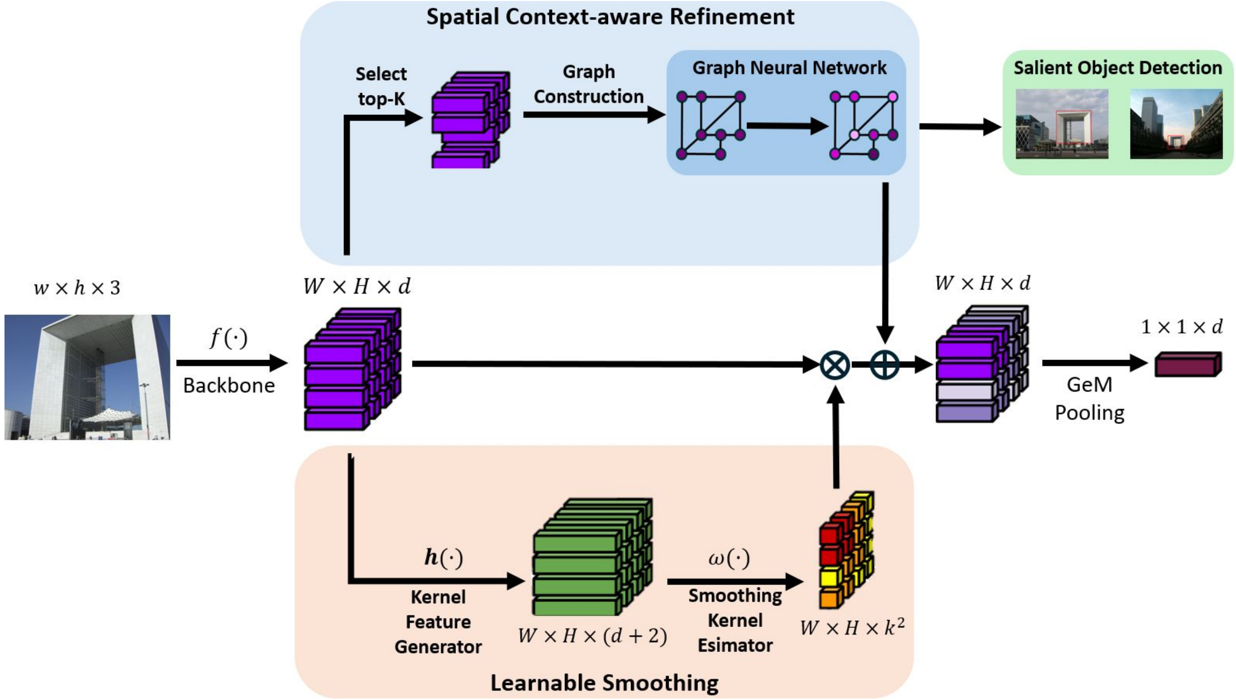
- SOD Results on ROxford & RParis

Summary

- Increase matching performance using global descriptor by providing spatial context of local features
- Learnable smoothing and Graph neural network for local & global spatial context extraction
- Analyze spatial context via localization performance



Plans & Schedule

- Jinhwan : Spatial Context-aware Refinement
- Kyu Beom : Learnable Smoothing

- Week 3-7. Survey
- Week 8. Mid-term
- Week 9. Install & baseline setup
- Week 10. Mid-term presentation
- Week 11-12. Implement our methods
- Week 13. Additional tuning for merging
- Week 14. Prepare final presentation
- Week 15. Final presentation