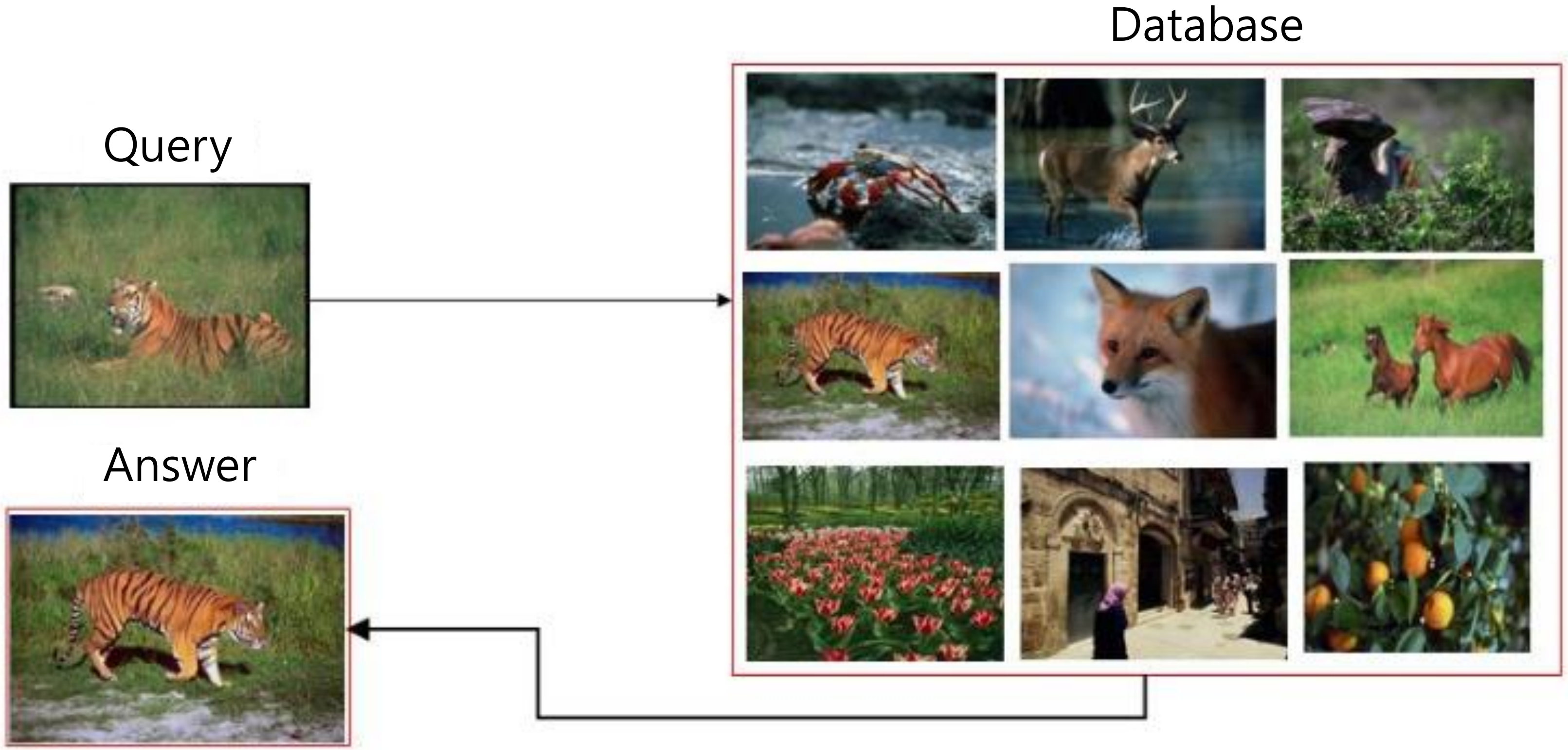
Regional Attention Based Deep Feature for Image Retrieval

Jaeyoon Kim (김 재 윤)

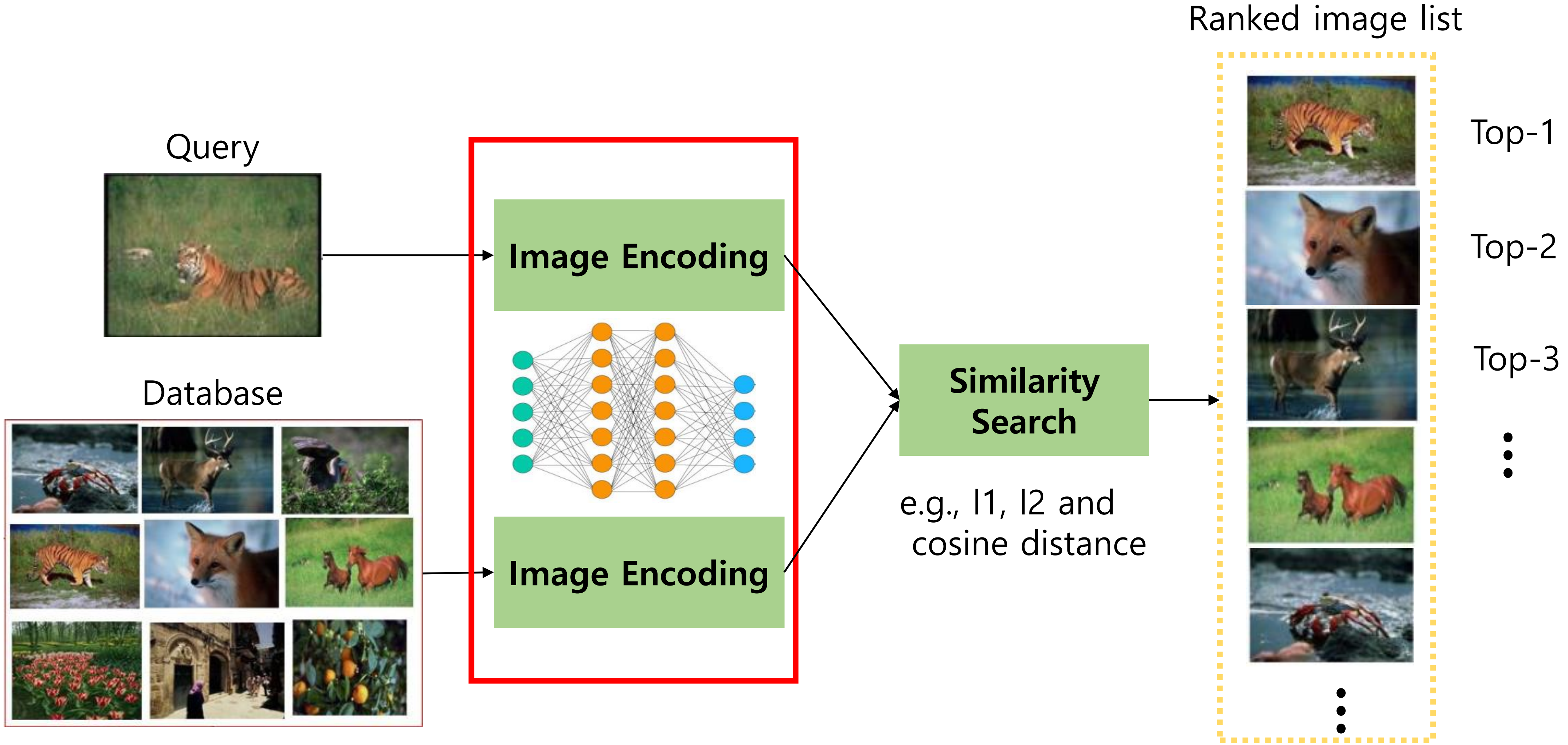
Advisor: Prof. Sung-Eui Yoon

Background – Image Retrieval (Object Retrieval)

- Given a query image, try to **find visually similar images** from an image database.

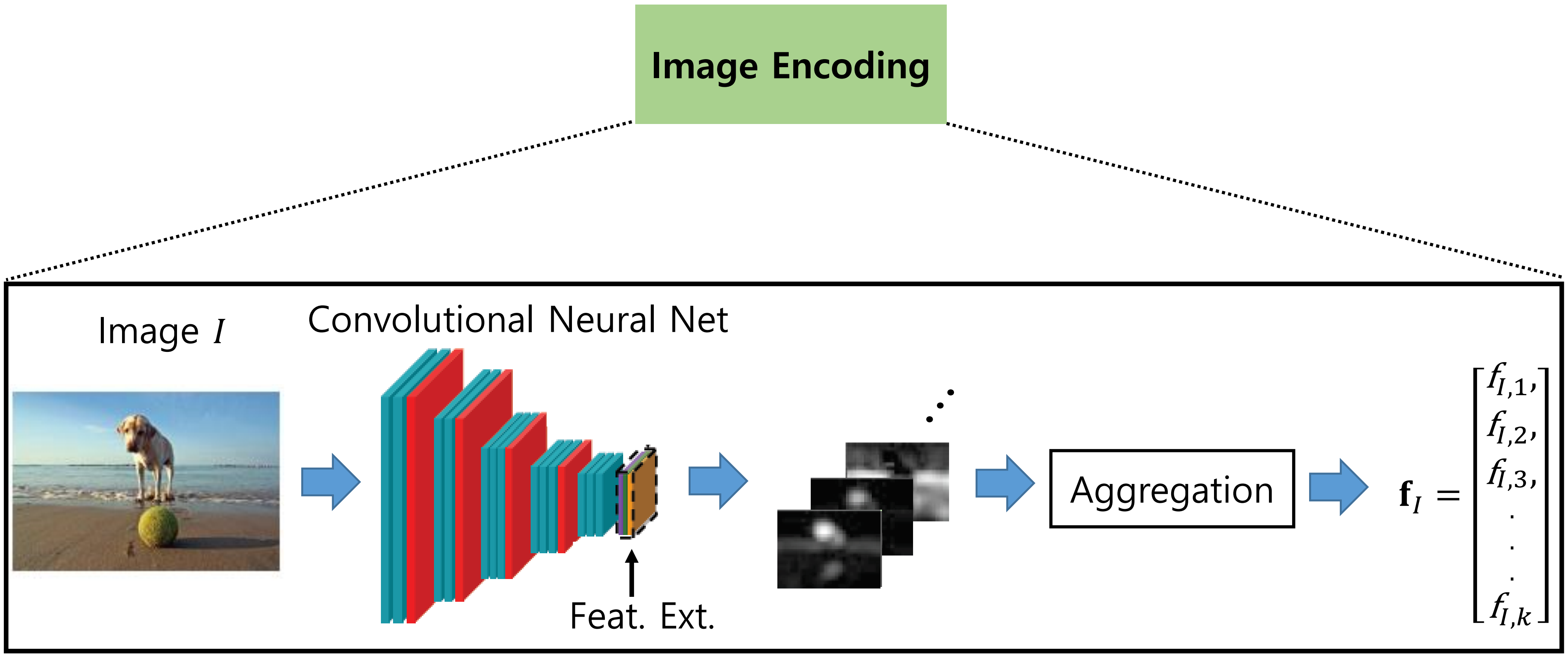


Background – General Pipeline of Image Retrieval



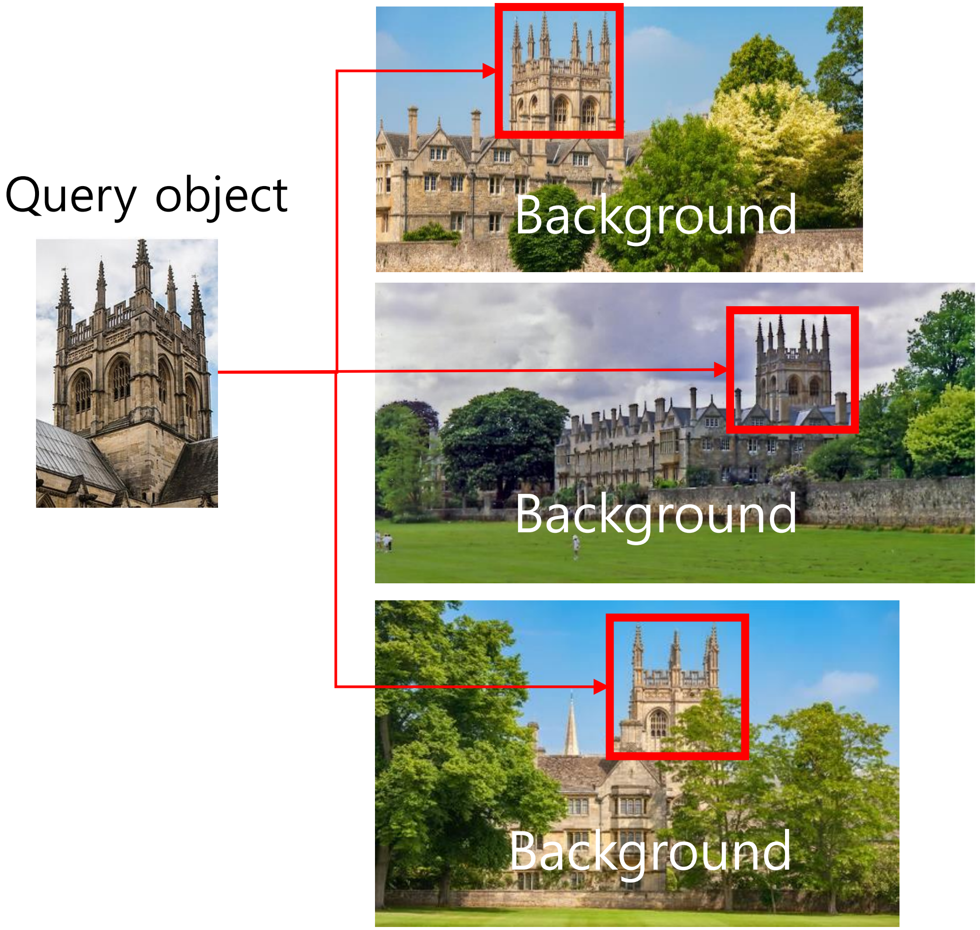
Background – Image Encoding (Image Embedding)

- Need to describe an image I into a **single feature vector $\mathbf{f}_I$**.



Problem – Common Challenges in IR

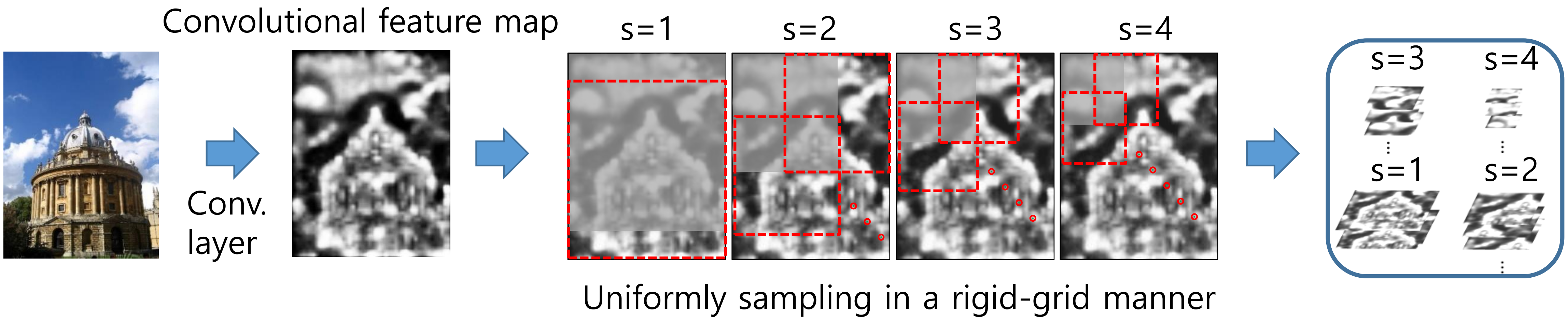
Background: Acted as a **distractor** while doing the aggregation



Object clutter: Where do we focus on?

Problem – R-MAC, Tolias *et al.*, 2015

- A commonly used technique for image retrieval.
- **Region-based aggregation** method with deep local features.



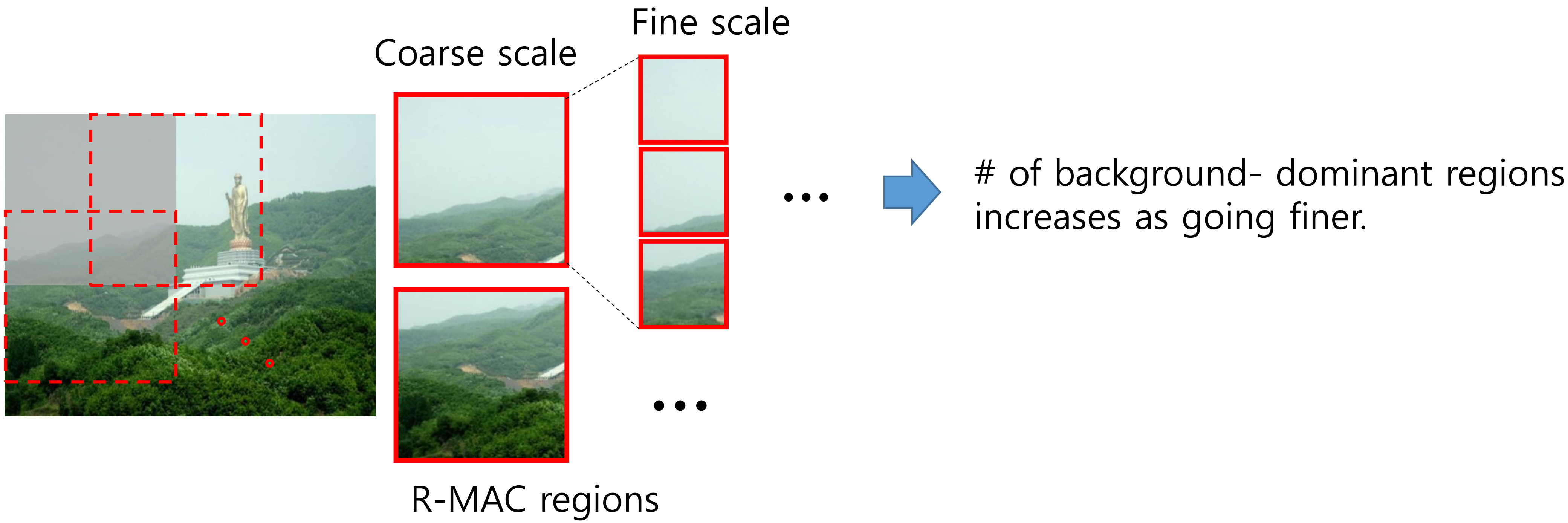
- **Vulnerable to backgrounds and object clutters** due to the uniform region sampling.

Tolias *et al.*, "Particular object retrieval with integral max-pooling of CNN activations", *CoRR, abs/1511.05879*, 2015.

Problem – R-MAC, Tolias *et al.*, 2015

- **A Conflict between below claim 1 and 2 in terms of region size**

1. **Vulnerability** becomes worse as small regions are sampled.



2. **Important to consider the small regions** in order to get a good result, especially in small object retrieval.

Problem – R-MAC, Tolias *et al.*, 2015

- Accuracy variation with different scales.
  - Verify the conflict is valid or not.

Oxford 5k(mAP)

Method	Scale (S)			
	S=3	S=4	S=5	S=6
R-MAC	69.9	**70.7**	70.1	69.0

No longer increase from the scale of 4 or more due to the background noise.

Objectives

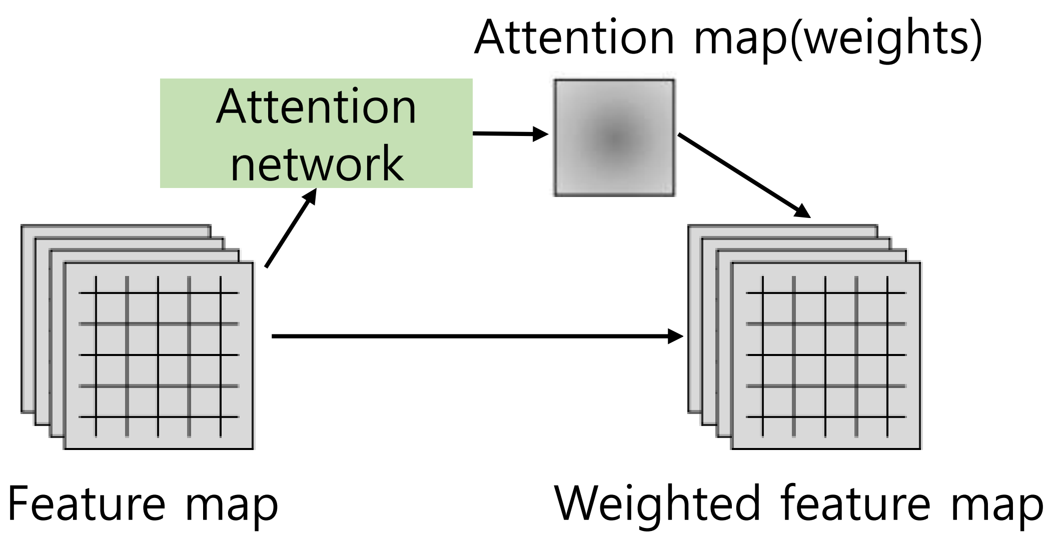
- Limitations of R-MAC are:
  - **Not to use a finer scale parameter** due to the background noise.
  - **Not to consider varying importance** among objects in object-cluttered images

- Goals are to:
  - Use our **regional attention network** for filtering the background noise.
  - Use our **context-awareness strategy** for considering varying importance.

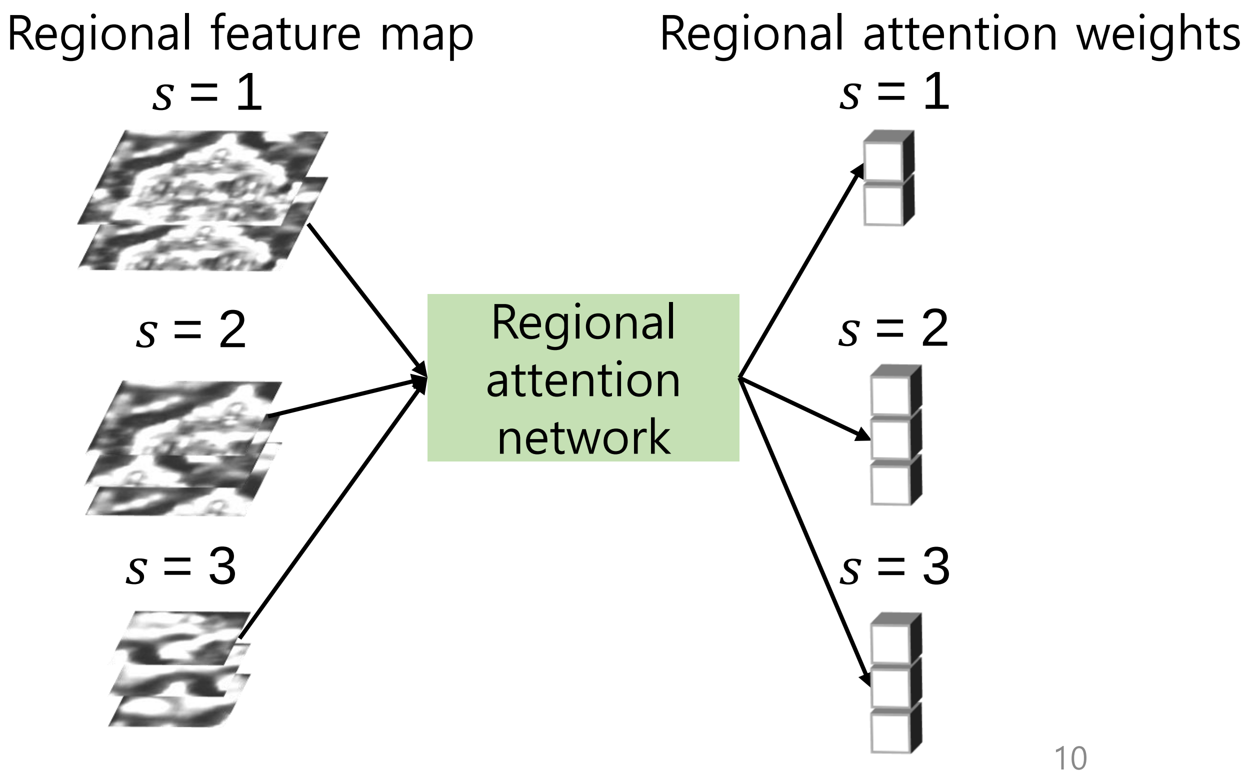
Our Approach – Regional Attention

- Propose regional attention network to filter the background noise.

- An existing pixel-based attention

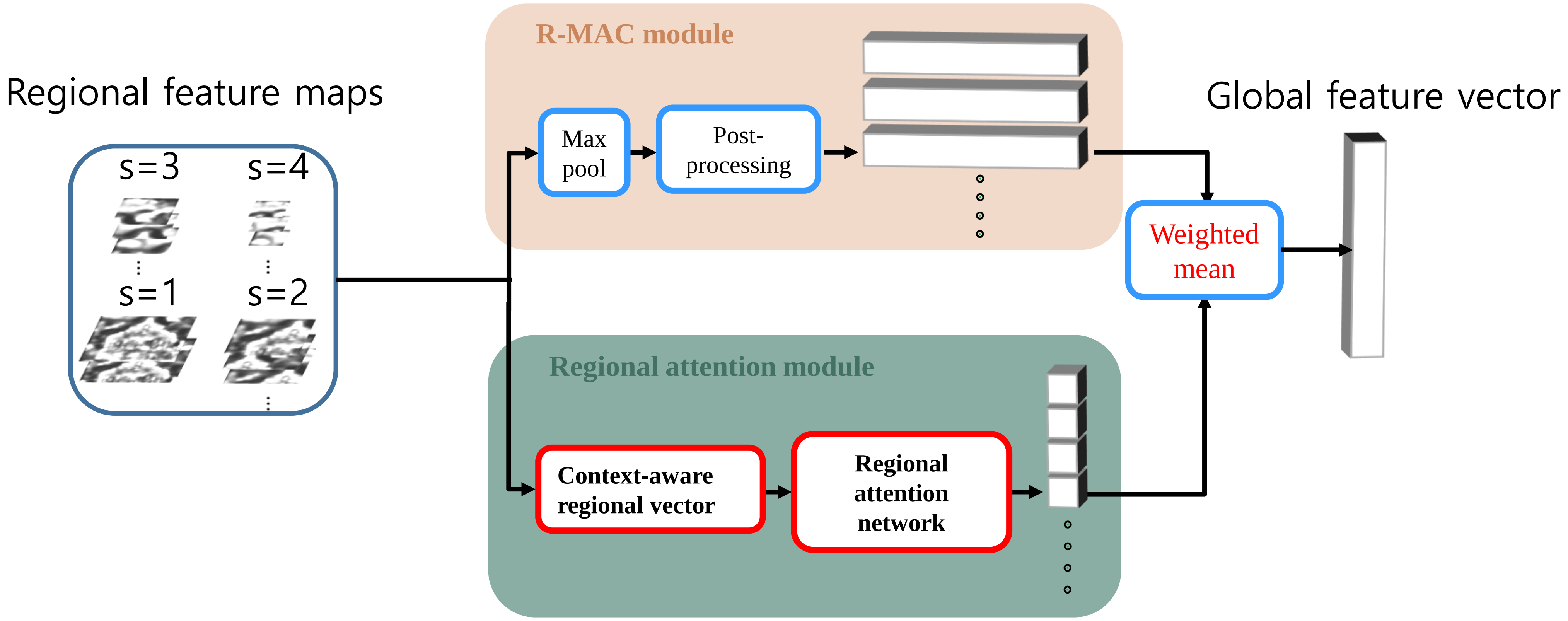


- Our region-based attention



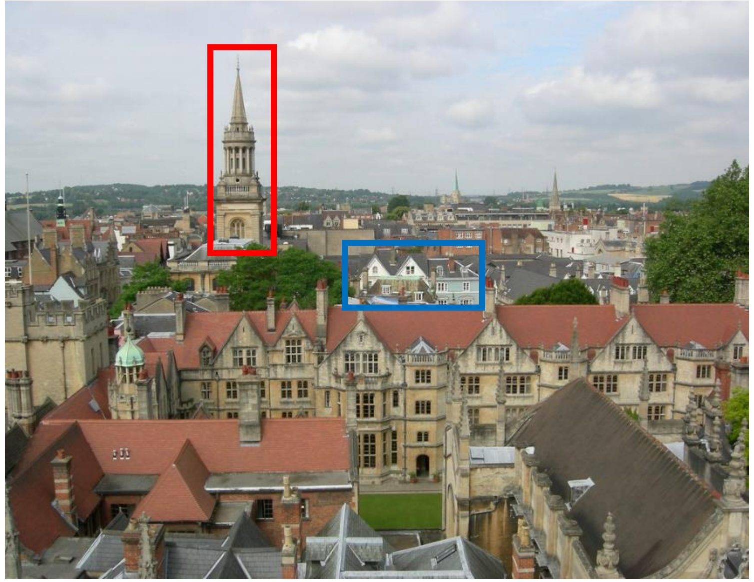
Our Approach – Regional Attention with R-MAC

- How our regional attention collaborates with R-MAC.



Our Approach – Context-Awareness

- Typically, people see an overall context of an image and then determine whether an object is salient or not [1].



vs

Which of objects is more important?
w/o context: ambiguity
w/ context: red one

- Consider a **global context** to get high-quality attention weights of local regions.

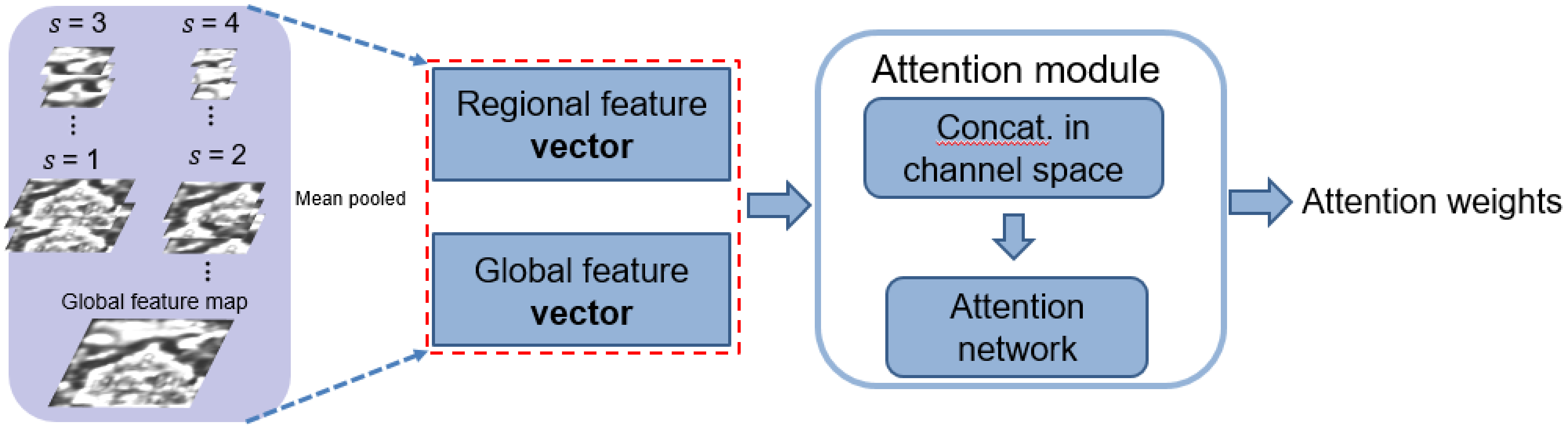
[1] Itti *et al.*, "Computational modelling of visual attention." Nature Reviews Neuroscience, 2001.

Our Approach – Regional Attention Architecture

- Use **regional** and **global** features to reflect the context-awareness
- Two linear layers and two non-linear layers

$$\Phi(\mathbf{k}) = \text{softplus}(\mathbf{W_c}\pi(\mathbf{k}) + \mathbf{b_c}),$$
$$\pi(\mathbf{k}) = \tanh(\mathbf{W_r k} + \mathbf{b_r}).$$

$\mathbf{k}$: Regional feature vector, $\Phi(\mathbf{k})$: Attention weight of $\mathbf{k}$



Our Approach – Ablation Study of Regional Attention

- Performance variation when our methods are added.

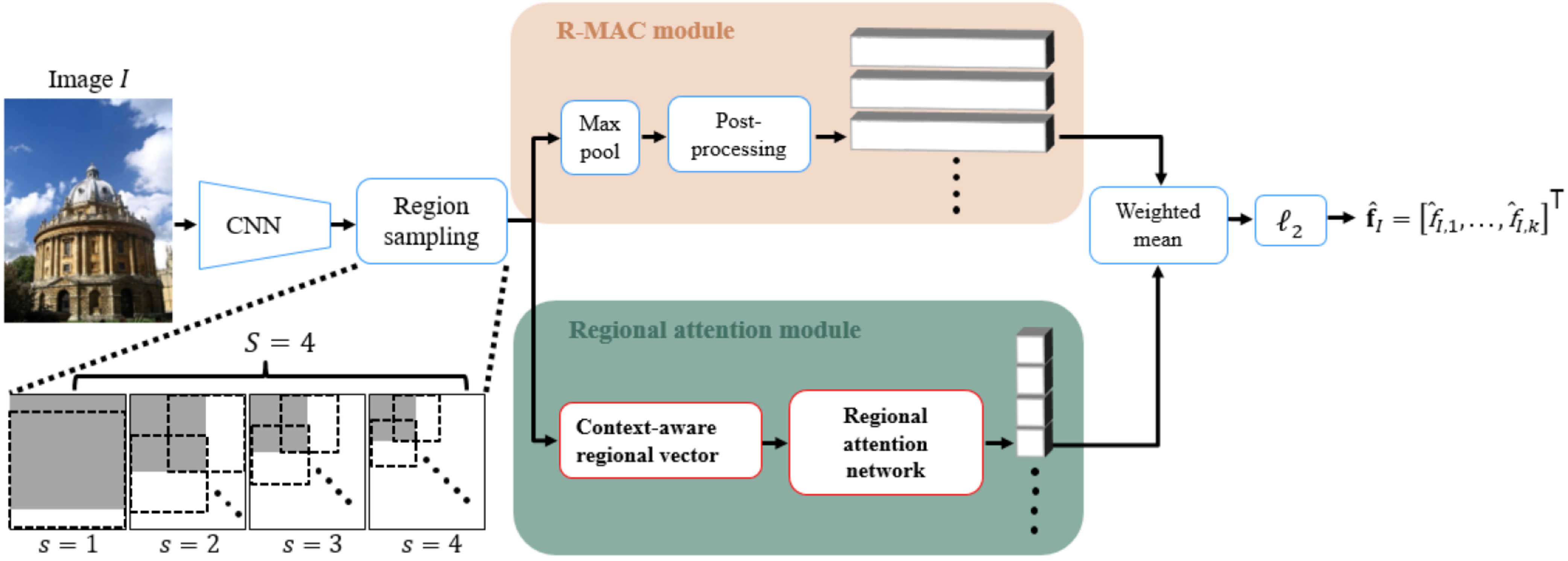
Measurement unit: mAP(mean Average Precision)

	Method	Oxford5k	Paris6k	Time (s)
	Baseline + PCA Landmark	70.1	85.4	0.095
Our methods	+ Regional attention	74.9	86.0	0.115
	+ Context awareness	76.8	87.5	0.123

<u>Our approach **consistently improved** when each module was added.</u>

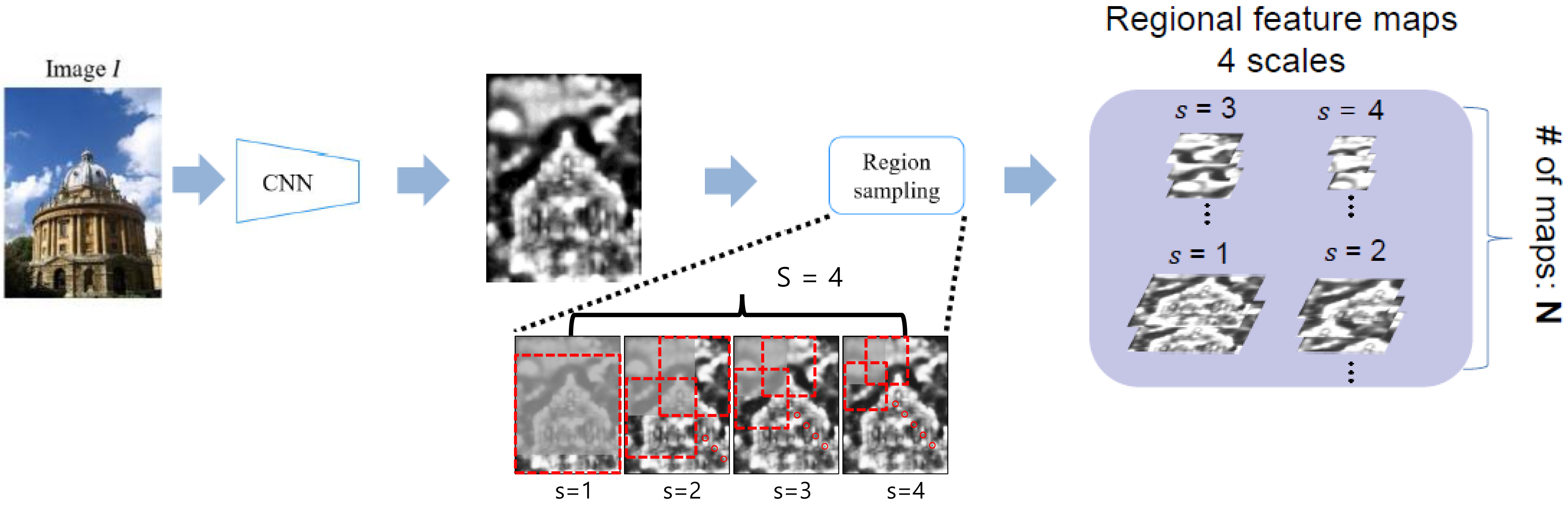
Our Approach – Image Encoding Pipeline

- Goal: Extract a global feature vector $\hat{\mathbf{f}}_I$ from an input image I



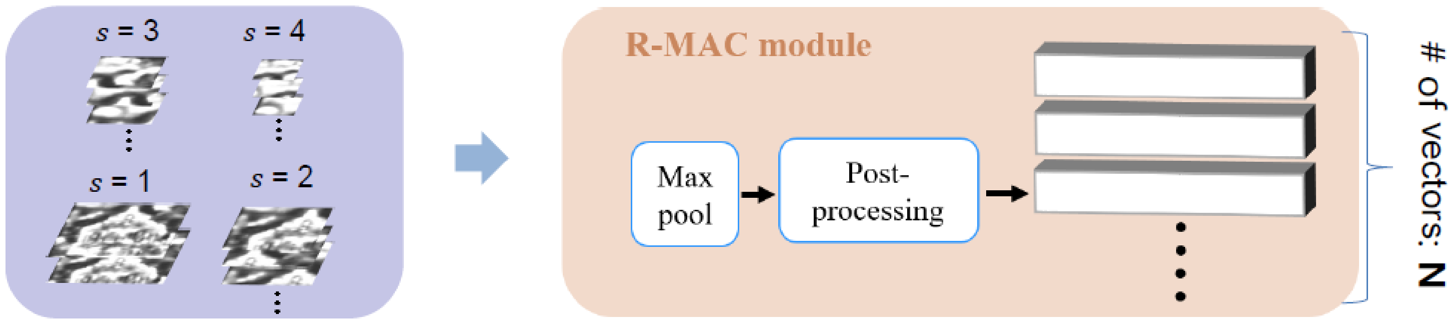
Our Approach – Image Encoding Pipeline

1. Extract a CNN feature map and sample regional feature maps in an R-MAC manner

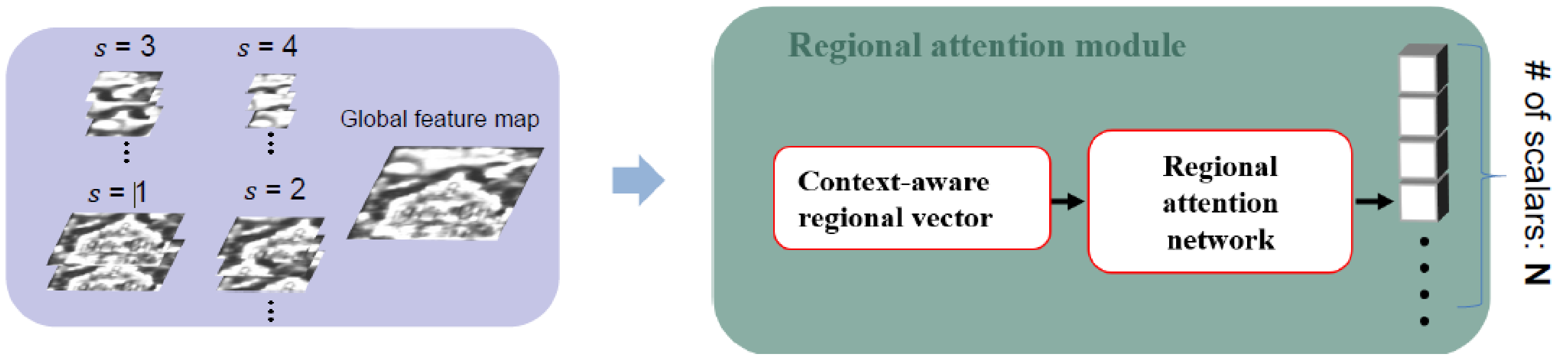


Our Approach – Image Encoding Pipeline

2.1 Produce R-MAC feature vectors with the regional feature maps

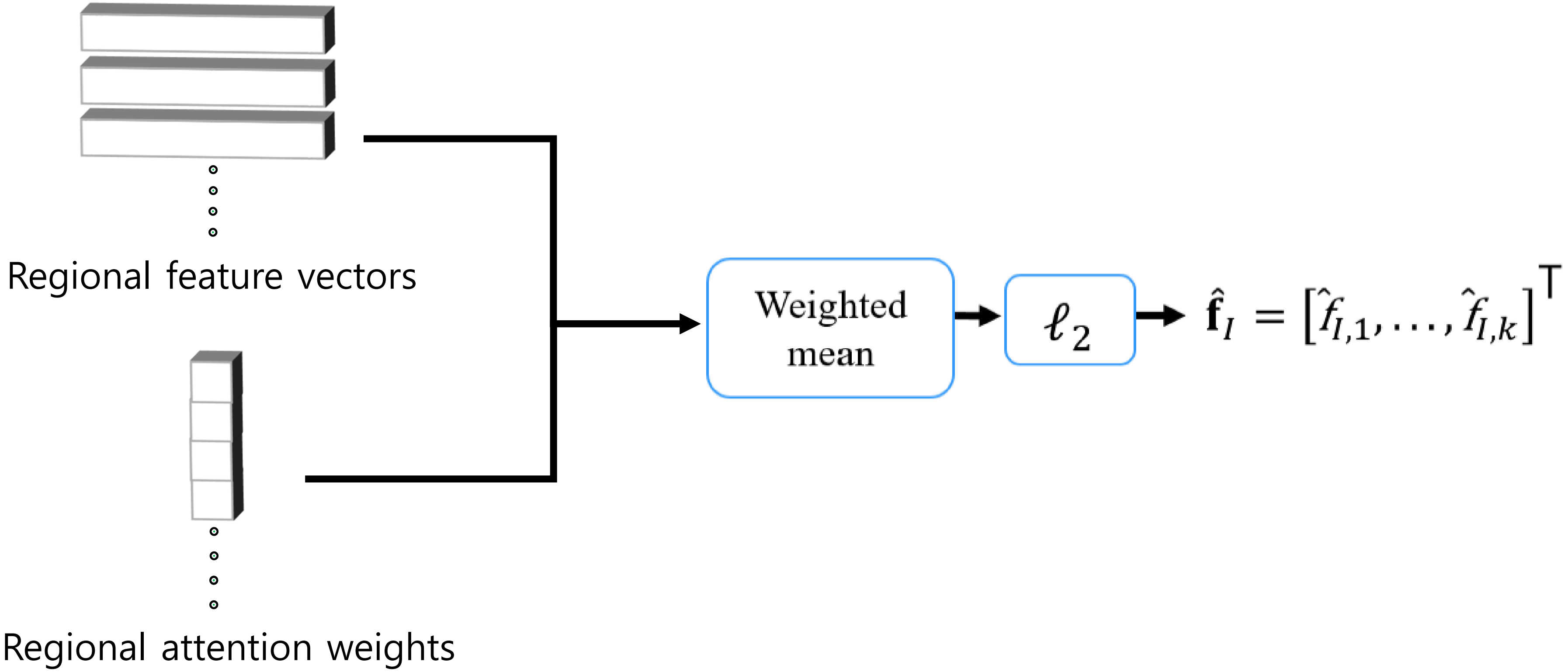


2.2 Calculate regional attention weights



Our Approach – Image Encoding Pipeline

3. Obtain a global feature vector, $\hat{\mathbf{f}}_I$, through combining R-MAC features with regional attention weights.



Our Approach – Summary of Main Contributions

- Propose **context-awareness** as well as **regional attention network**.
- Improve **R-MAC** with a large gap of accuracy.
- Achieve a new state-of-the-art performance in IR.

Result & Analysis

- Experiment
  - Training dataset : ImageNet 1M – 1000 classes
  - Benchmark datasets
    - *Oxford 5k*: landmark (building) images in Oxford
    - *Oxford 105k*: Oxford 5k + 100k distractor images
    - *Paris 6k*: landmark (building) images in Paris
    - *Paris 106k*: Paris 6k + 100k distractor images

Result & Analysis

- Comparison with the state-of-the-arts

Our approach consistently set a state-of-the-art accuracy for **each 4 dataset with large gaps**!

Measurement unit: mAP(mean Average Precision)

	Method	Dim.	Oxford5k	Paris6k	Oxford105k	Paris106k
Resnet101	SDCF [8]	2048	69.1	81.7	65.4	74.3
	CroW [14]	2048	68.7	82.8	62.7	75.1
	R-MAC [27]	2048	70.1	85.4	66.9	80.8
	CAM [13]	2048	69.9	84.3	64.3	77.1
	Ours	2048	**76.8**	**87.5**	**73.6**	**82.5**
	Query expansion (QE)					
Resnet101	SDCF+QE [8]	2048	68.5	84.9	66.8	79.4
	CroW+QE [14]	2048	69.5	85.1	66.7	79.9
	R-MAC+QE [27]	2048	73.8	86.4	71.8	82.6
	CAM+QE [13]	2048	71.3	86.1	68.7	80.8
	Ours+QE	2048	**81.8**	**89.3**	**80.4**	**85.4**

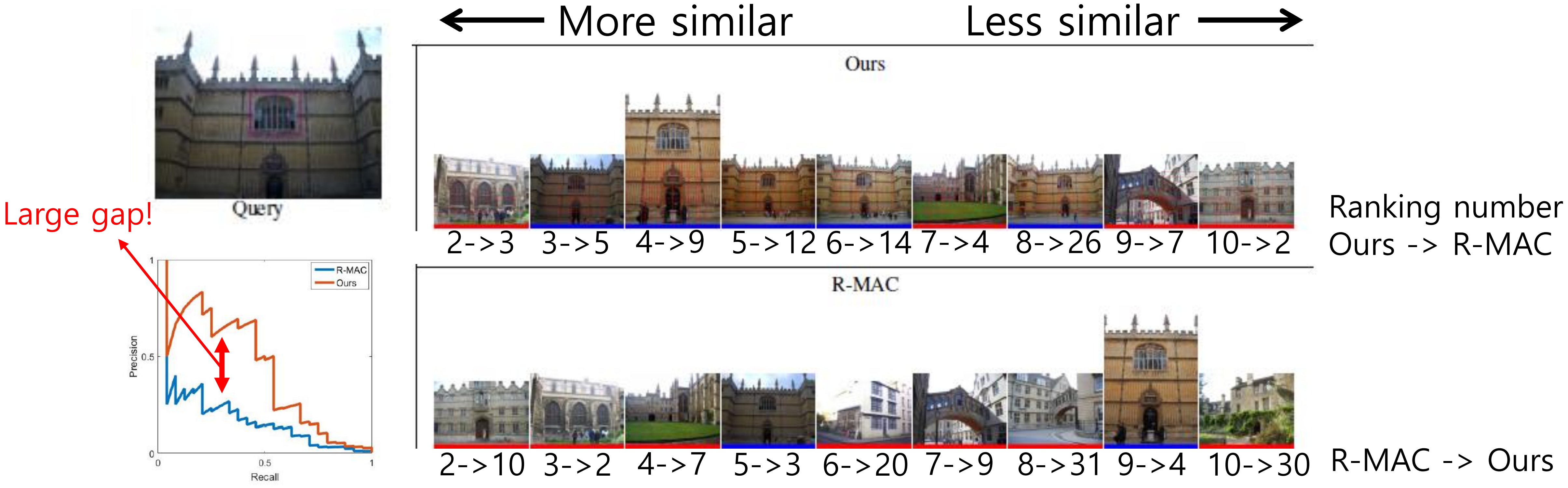
Query expansion: commonly used in IR as an existing material

Average improvement gap: 4.1mAP
Other methods: 2~3 mAP

➡ The larger gap with query expansion!

Result & Analysis - Qualitative Comparison

- Show one example (query) where ours outperforms R-MAC
- Note that ours surpasses R-MAC in 54 queries out of 55 queries.

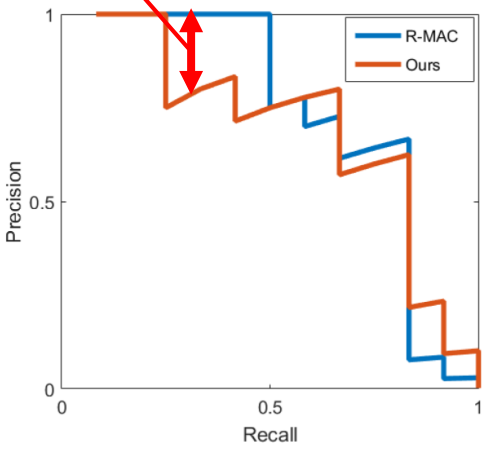


Result & Analysis - Qualitative Comparison

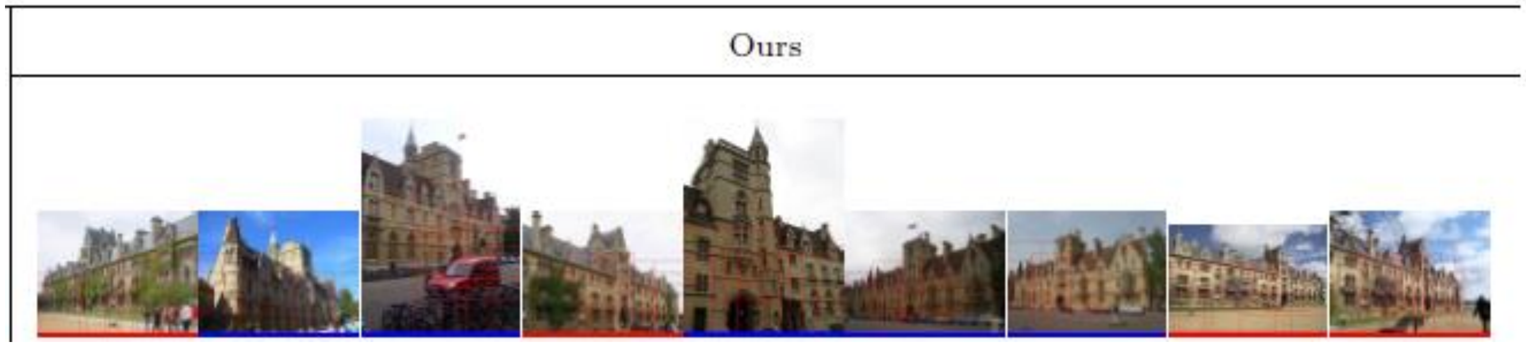
- Qualitative results – comparison with R-MAC
  - **Only single case** where R-MAC outperforms ours

Query

Small gap!

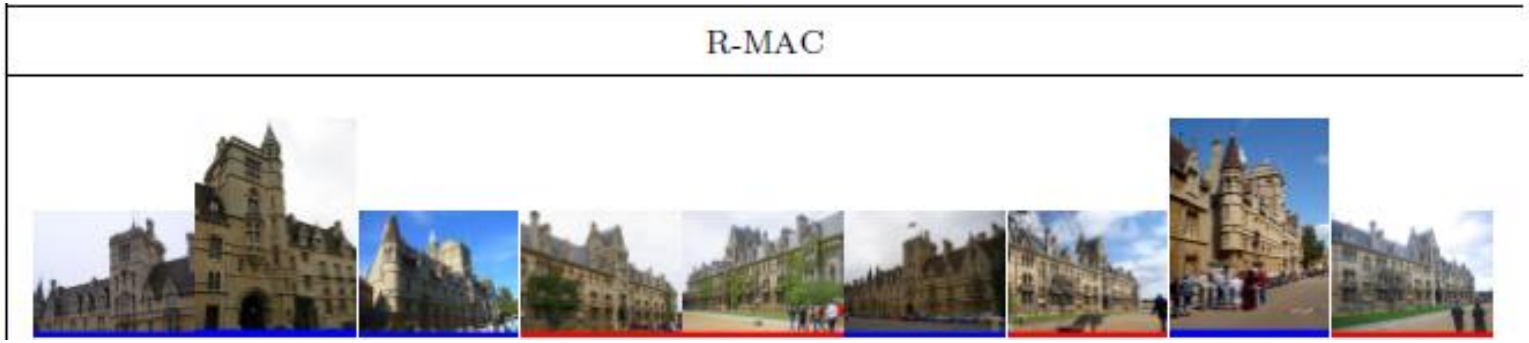


More similar   Less similar



4->8 5->6 6->3 7->7 8->5 9->9 10->15 11->23 12->10

Ranking number
Ours -> R-MAC



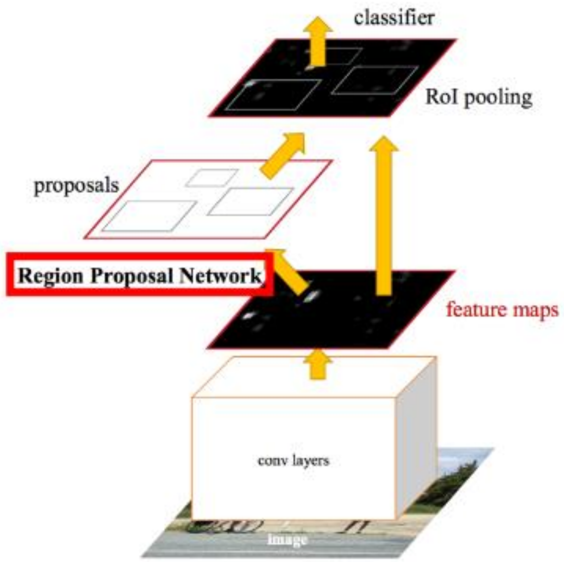
4->3 5->8 6->5 7->7 8->4 9->9 10->12 11->15 12->13

R-MAC -> Ours

Result & Analysis

- Ablation Study – Region Proposal Network(**RPN**)
  - RPN has been employed in various computer vision tasks.

Our approach can cooperate with **another region-sampling method(RPN)** as well as R-MAC.



Measurement unit: mAP(mean Average Precision)

Method	Oxford5k	Paris6k
RPN + PCA Landmark	64.7	75.5
+ Regional attention	66.6	75.8
+ Context awareness	67.9	76.4

Conclusion

- Introduced regional attention network to handle the background noise and object clutter.

- Set the state-of-the-art **search accuracy**, especially in query expansion by:
  - Combining R-MAC with our regional attention network.
  - Utilizing a context-awareness strategy for regional attention network.

- Showed a **generality** of our regional attention network by cooperating with **region proposal network(RPN)**.

Future Work

- Apply our-method-equipped RPN into other tasks of computer vision
- Consider the **fine-tuning using metric learning**

Publication

- **J. Kim**, S. Yoon, “Regional Attention Based Deep Feature for Image Retrieval”
  - Published in BMVC, 2018

- **J. Kim**, S. Um, D. Min, “Fast 2D Complex Gabor Filter with Kernel Decomposition”
  - Published in IEEE TIP, April, 2018

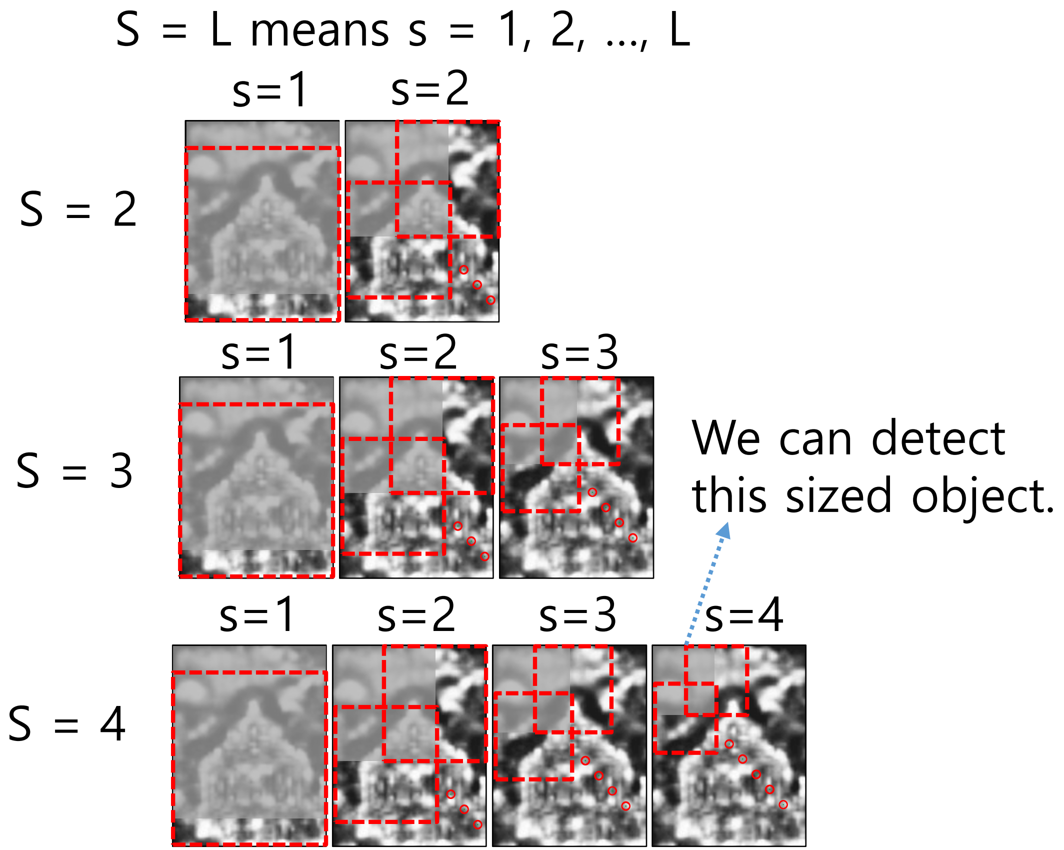
Thank you for listening!
Q&A

Acknowledgements

Advisor Sung-Eui Yoon & SGVR members

Problem – R-MAC, Tolias *et al.*, 2015

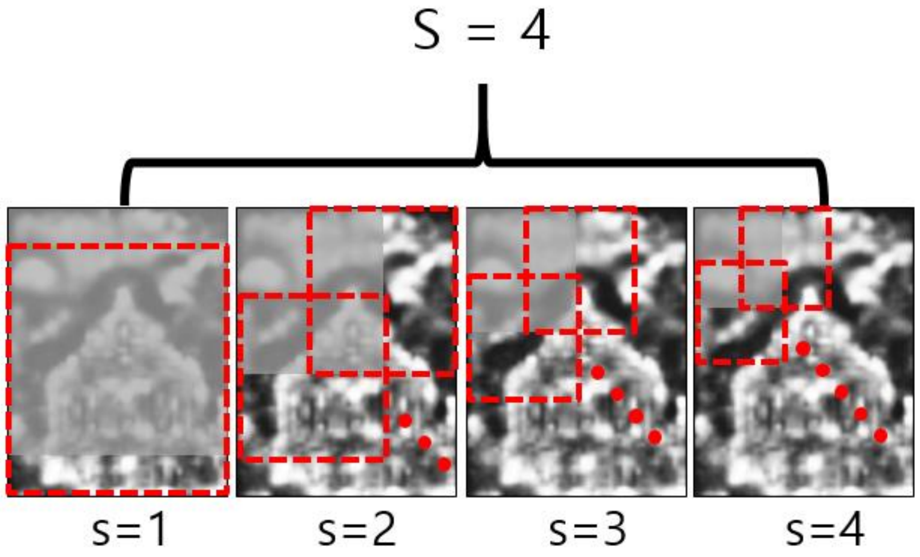
- The bigger the scale we choose, the smaller objects can be detected.



Result & Analysis

- Ablation Study – R-MAC vs ours in terms of scale
  - Our approach works robustly on **finer scale**.

Our approach **can see more details** of an image thanks to finer scale



Measurement unit: mAP(mean Average Precision)

Method	Scale (S)			
	S=3	S=4	S=5	S=6
R-MAC	69.9	**70.7**	70.1	69.0
Ours	75.1	76.7	**76.8**	76.4