최근 딥러닝 기반 이미지 검색 기술에 대한 소개

Jaeyoon Kim
(김재윤)

Outline

- Learning-based approaches
- Descriptor whitening
- Benchmarks (training and test data)
- Post-processing on online time

Most of this presentation materials was built upon Tolias's.

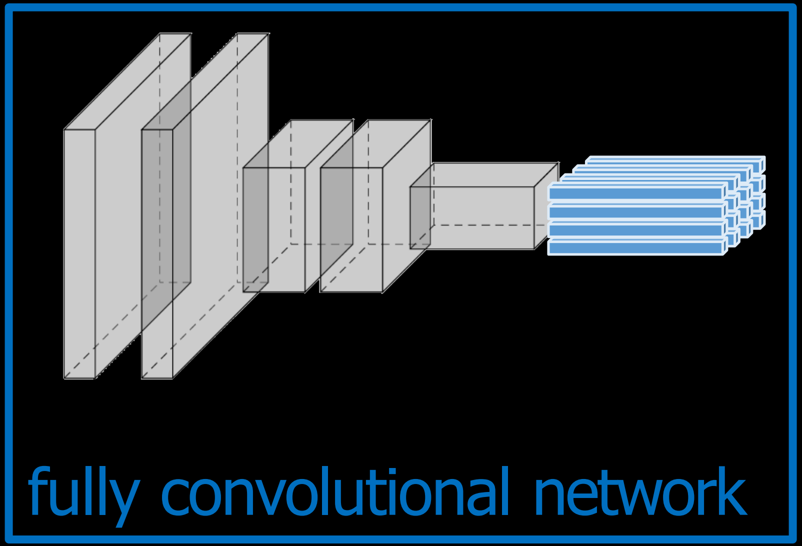
Learning-based methods

Global descriptor

$X = f($  $) \in \mathbb{R}^d$

- Instance search reduces to similarity search in d-dimensional space
- Compatible with efficient nearest neighbor techniques

Global descriptors with CNNs

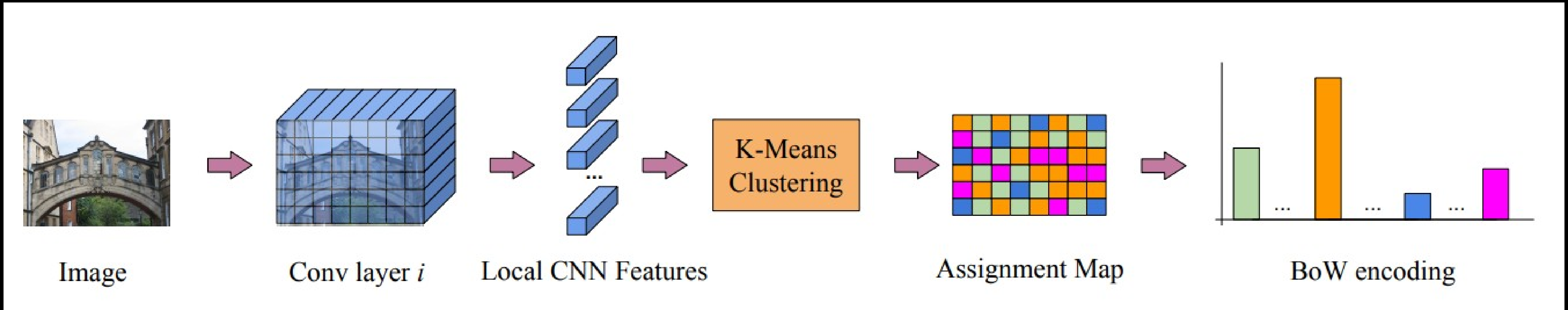


embedding & aggregation

descriptor:

$$X \propto \sum_{\mathbf{x} \in \mathcal{X}} g(\mathbf{x})$$

BoW with CNN features



- Used with pre-trained features and hard assignment
- Soft assignment needed for training

[Mohedano et al. ICMR'16]

Sum pooling – SPoC descriptor

- Descriptor

$$X \propto \sum_{\mathbf{x} \in \mathcal{X}} \mathbf{x}$$

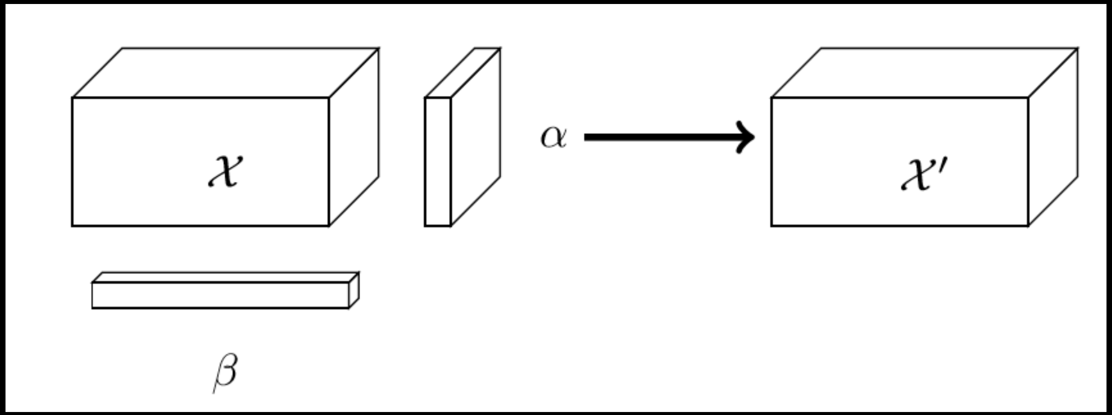
- Pair-wise similarity

$$X^\top Y \propto \sum_{\mathbf{x} \in \mathcal{X}} \sum_{\mathbf{y} \in \mathcal{Y}} \mathbf{x}^\top \mathbf{y}$$

- Simple but works
  → discriminative power of CNN activations

[Babenko & Lempitsky, ICCV'15]

Weighted sum pooling – CroW descriptor

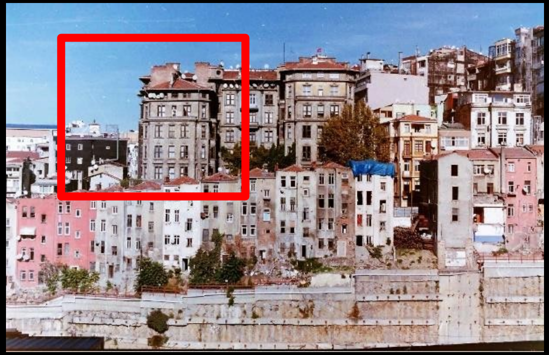


α: weight based on L2 norm of local descriptors
β: inverted-document-frequency weight

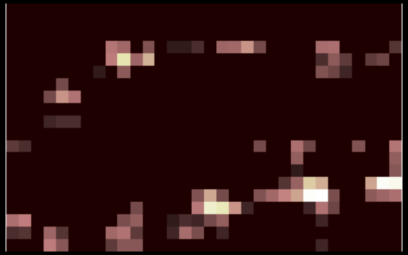
example of α

[Kalantidis et al., ECCV’16]

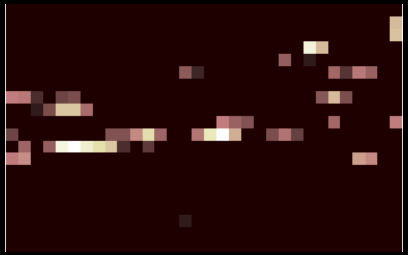
Max pooling – MAC descriptor



Input image

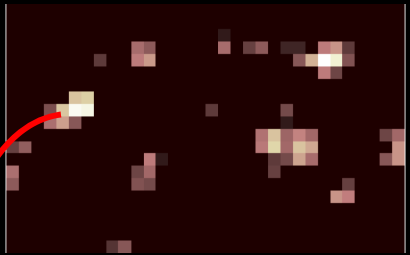


conv_5 filter 1



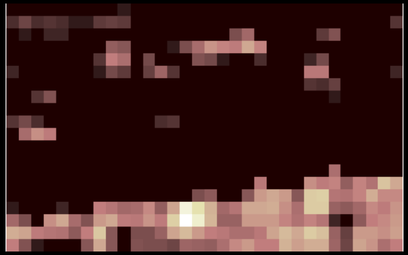
conv_5 filter 2

….



conv_5 filter i

….



conv_5 filter K

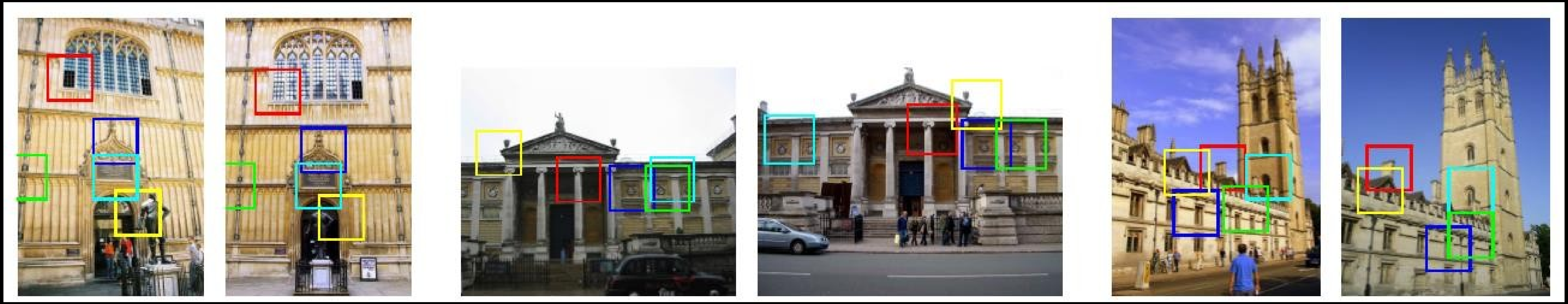
maximum activation

$$\text{MAC} = [f_1, \ldots, f_i, \ldots, f_K]$$

[Razavian et al., MTA'16] [Tolias et al., ICLR'16]

Max pooling – MAC descriptor

pair 1 pair 2 pair 3



regions for top matching components
different color per component

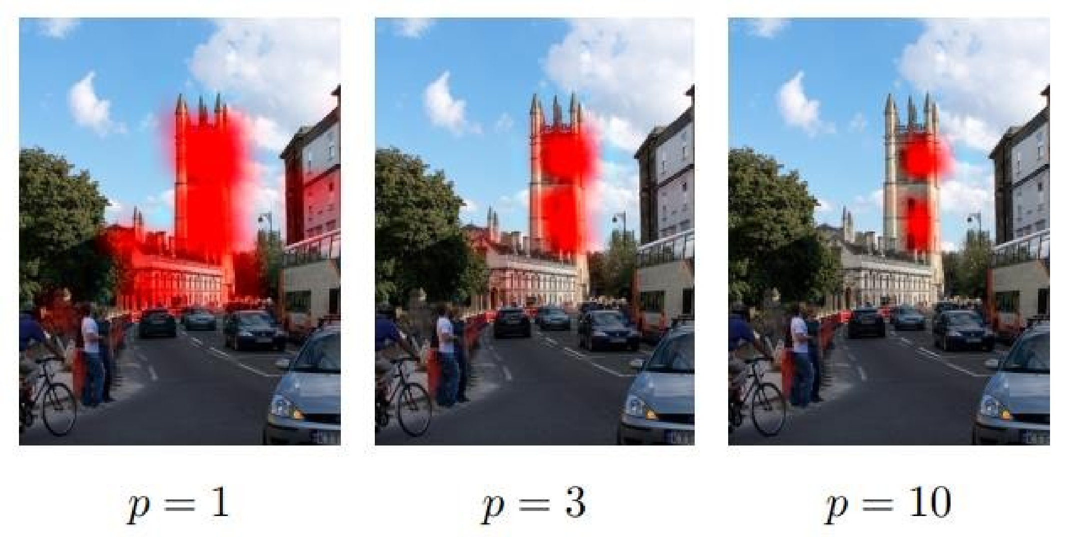
[Razavian et al., MTA'16] [Tolias et al., ICLR'16]

Generalized mean pooling – GeM descriptor

$$X \propto \left(\frac{1}{|\mathcal{X}|} \sum_{\mathbf{x} \in \mathcal{X}} \mathbf{x}^p \right)^{\frac{1}{p}}$$

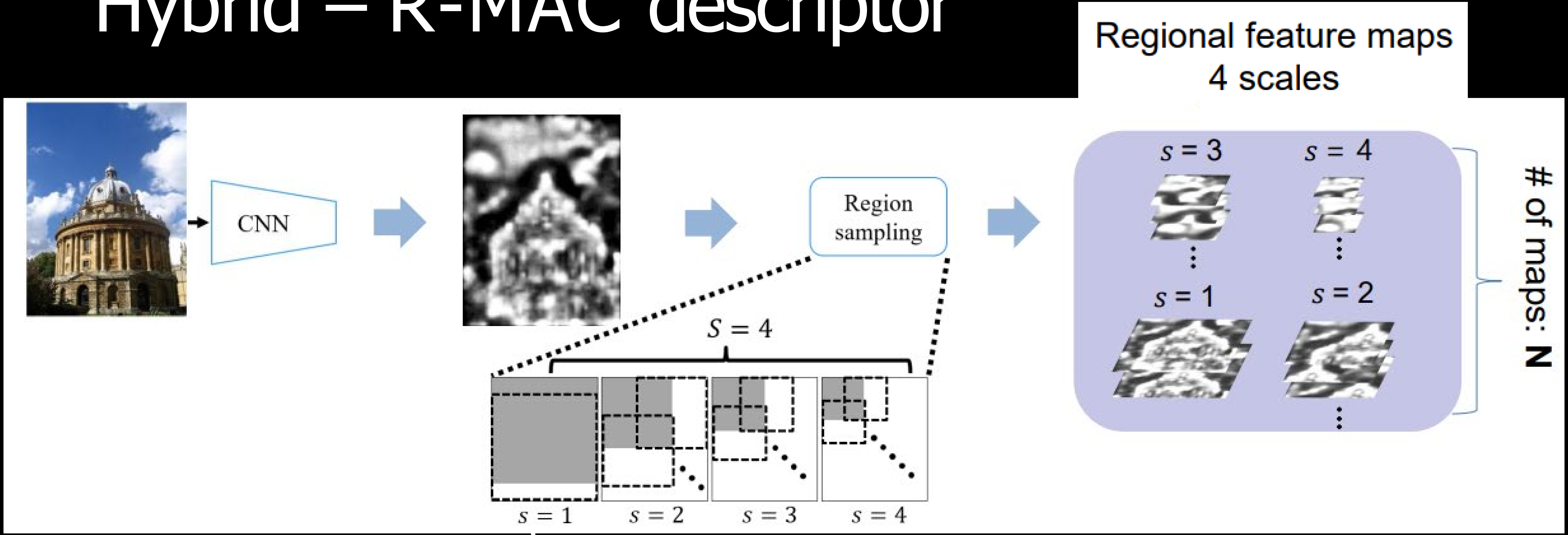
where $\mathbf{x}^p$ is element-wise power

$p \rightarrow \infty$ max pool (MAC)
$p = 1$ avg pool (SPoC)



[Radenovic et al., PAMI'19]

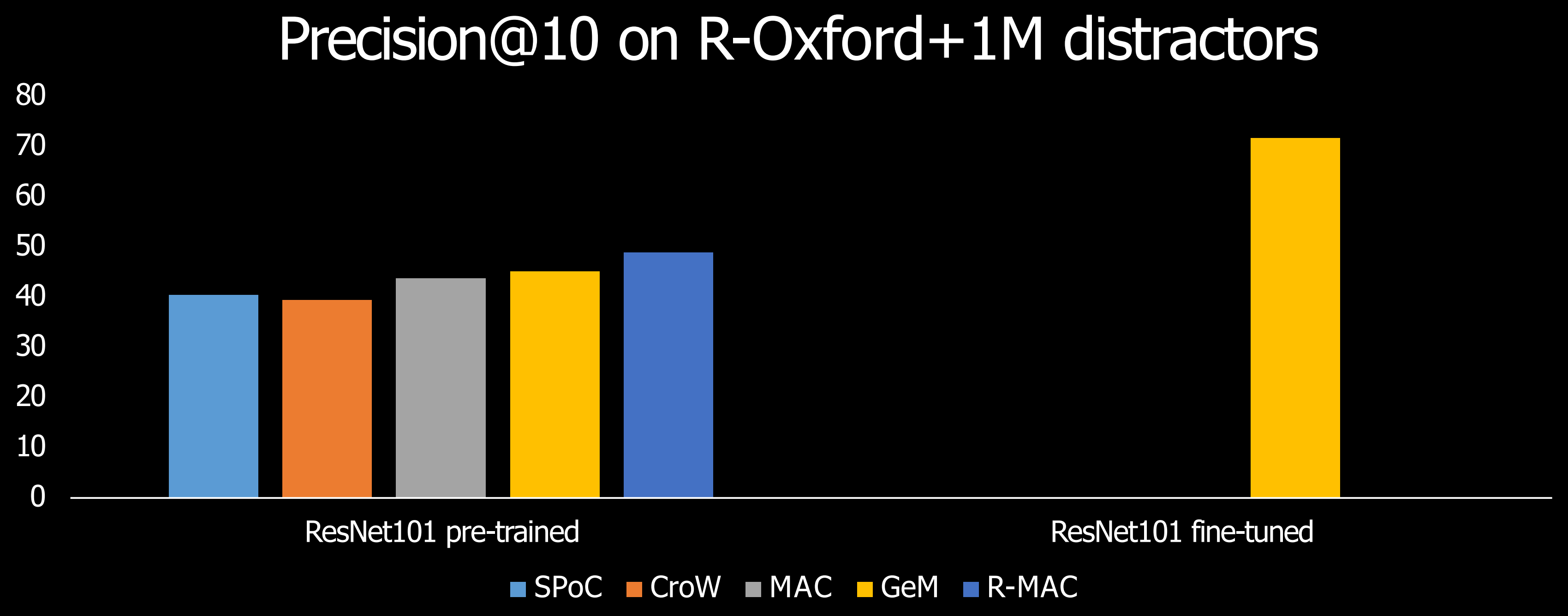
Hybrid – R-MAC descriptor



- Sum aggregate

[Tolias et al., ICLR’16]

Performance comparison



Fine-tuning improvement for GeM: +26.6%

Descriptor whitening

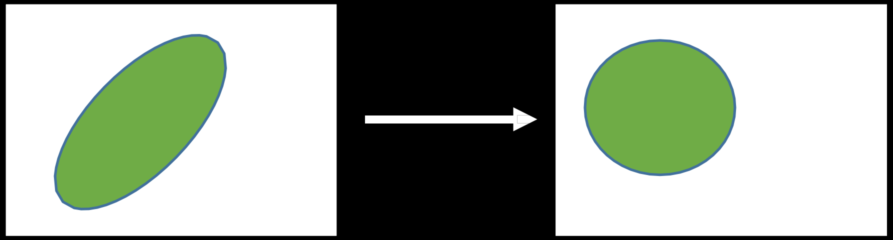
Descriptor processing with PCA

$$\hat{\mathbf{x}} = P^{\top}(\mathbf{x} - \mu)$$

$P \in \mathbb{R}^{d \times d}$ eigen-vectors as columns

$\mu \in \mathbb{R}^{d}$ mean vector glo

$\mathbf{x} \in \mathrm{R}^{d}$ bal descriptor



[Jegou & Chum, ECCV’12]

Post-processing with whitening

$\mathcal{X} = f($  $)$

$\mathcal{X} = \{\mathbf{x}\}, \ \mathbf{x} \in \mathbb{R}^d$

process & aggregate

post-processing (whitening)

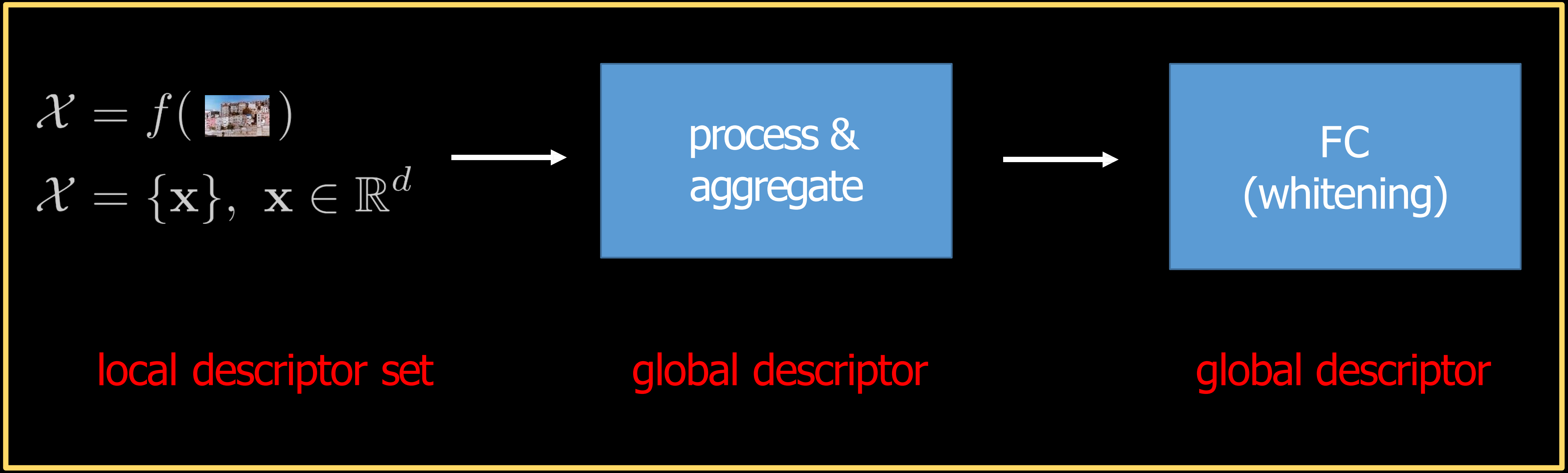
local descriptor set

global descriptor

global descriptor

learned end-to-end

Post-processing with whitening



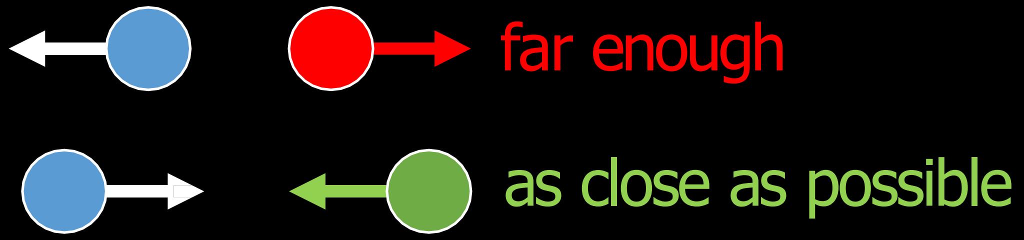
learned end-to-end

https://github.com/filipradenovic/cnnimageretrieval-pytorch

Training loss

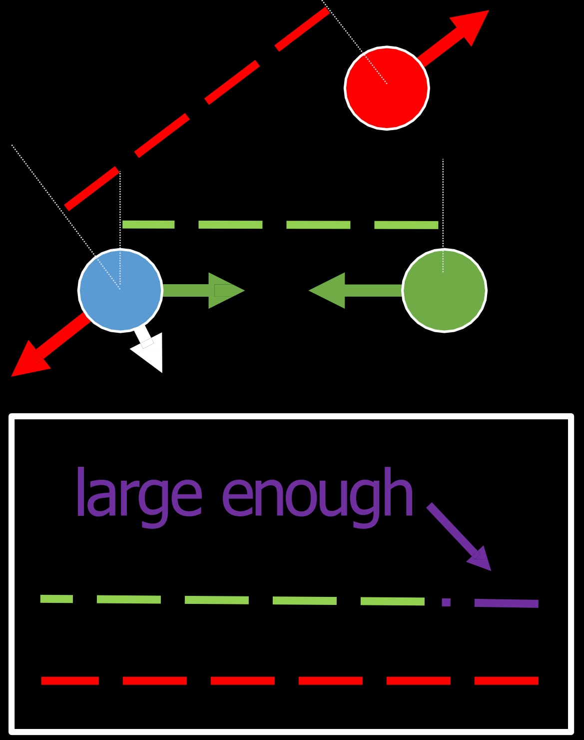
Loss functions for metric learning

Contrastive loss

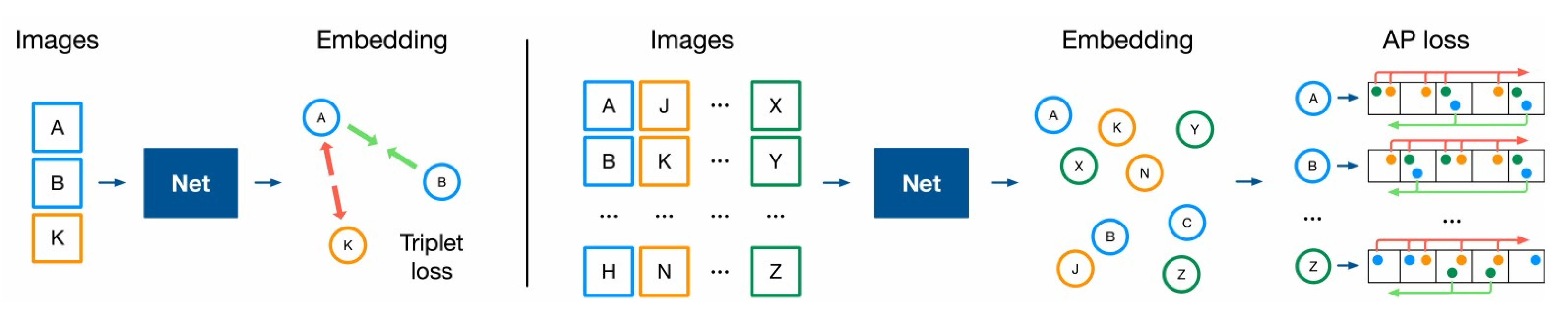


- Sampling from discrete class labels
  - problem: large intra-class variability
- Need automatic ways for pair-wise labels

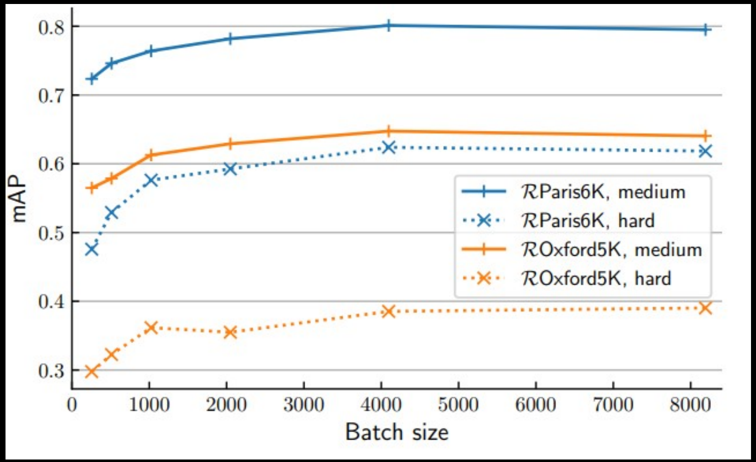
Triplet loss



Average precision loss



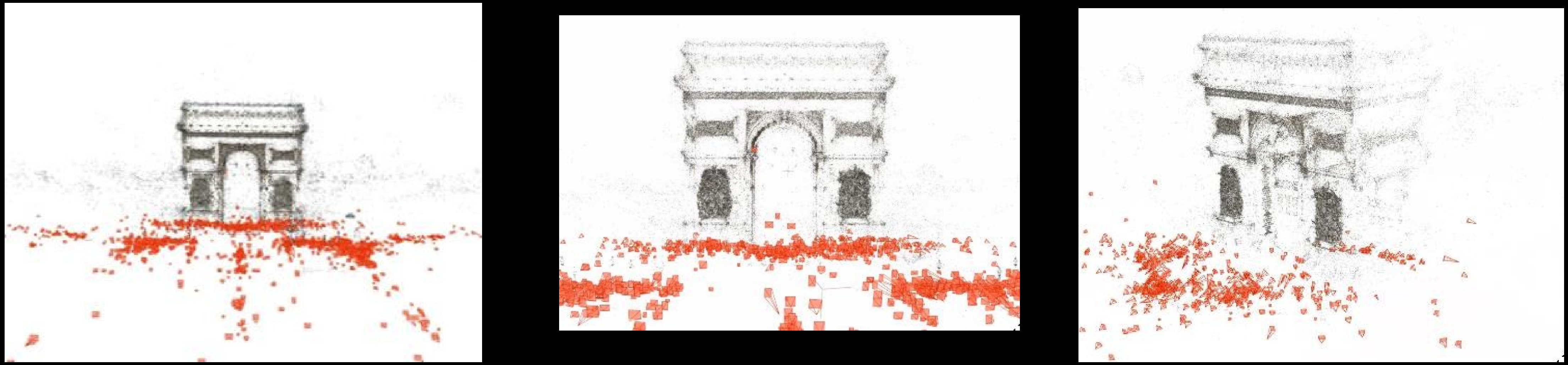
The larger the batch the better
→ no need to sample



[Revaud et al., ICCV'19]

Training data

Training data from SfM

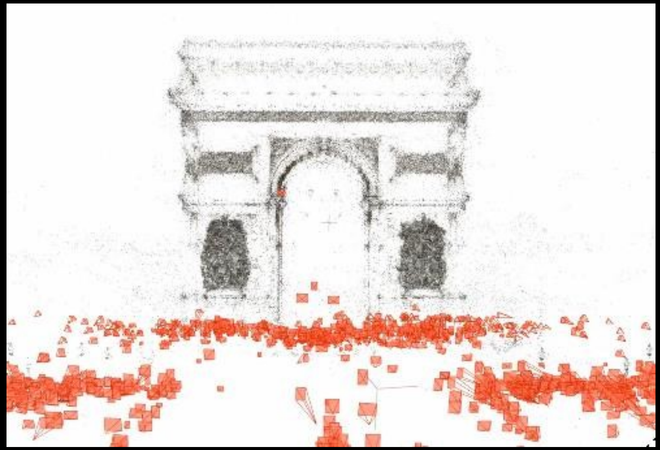
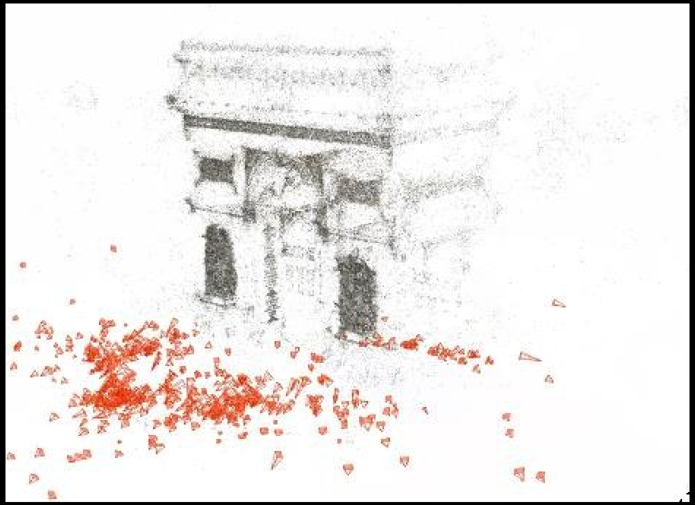


7.4M images → 713 training 3D models

[Schonberger et al. CVPR’15]
[Radenovic et al. CVPR’16]

Training data from SfM

camera orientation known
number of inliers known

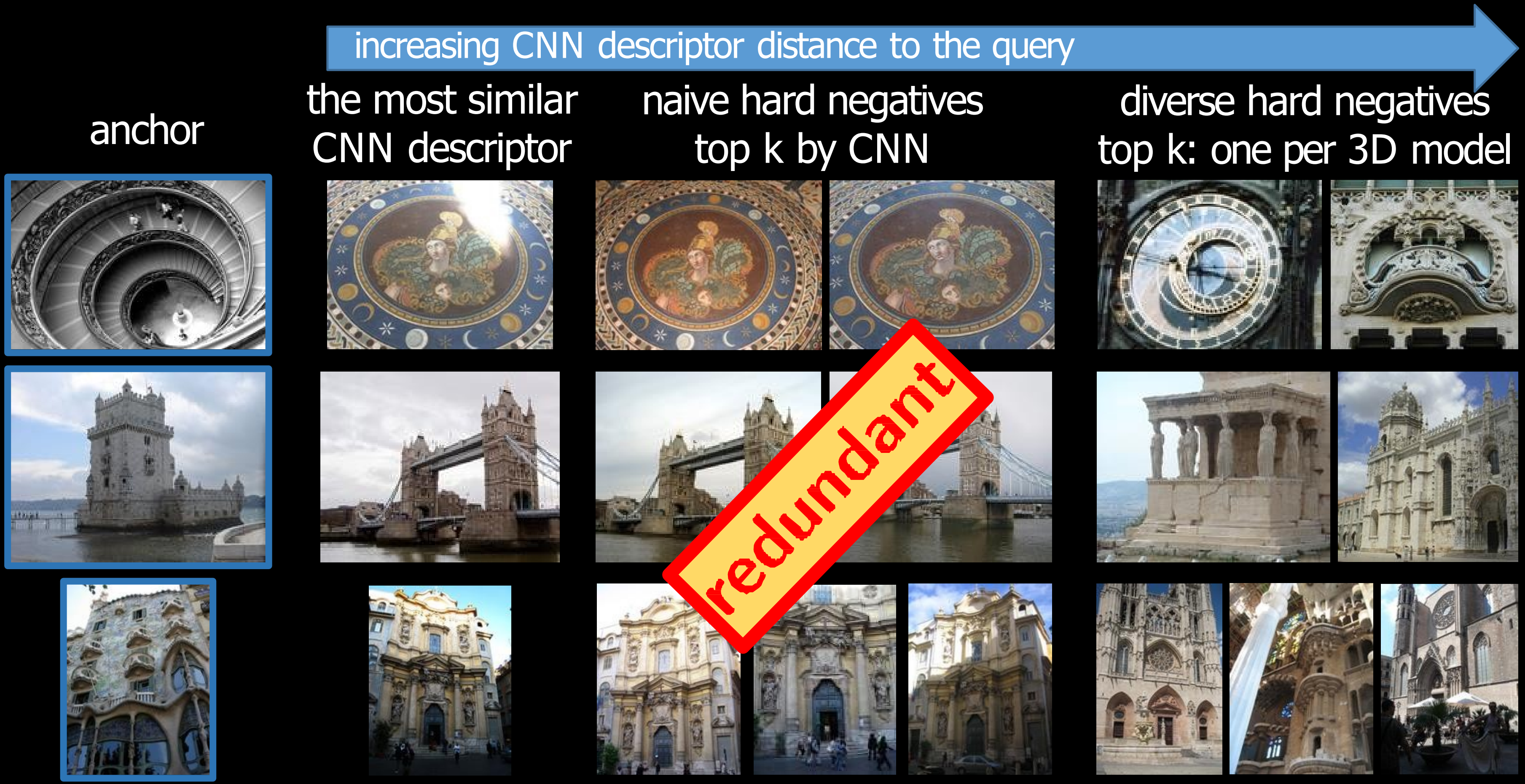




7.4M images → 713 training 3D models

[Schonberger et al. CVPR’15]
[Radenovic et al. CVPR’16]

Training data from SfM: hard negatives

Negative examples: images from different 3D models than the query
Hard negatives: closest negative examples to the query



[Radenovic et al. PAMI'19]

Training data from SfM: hard positives

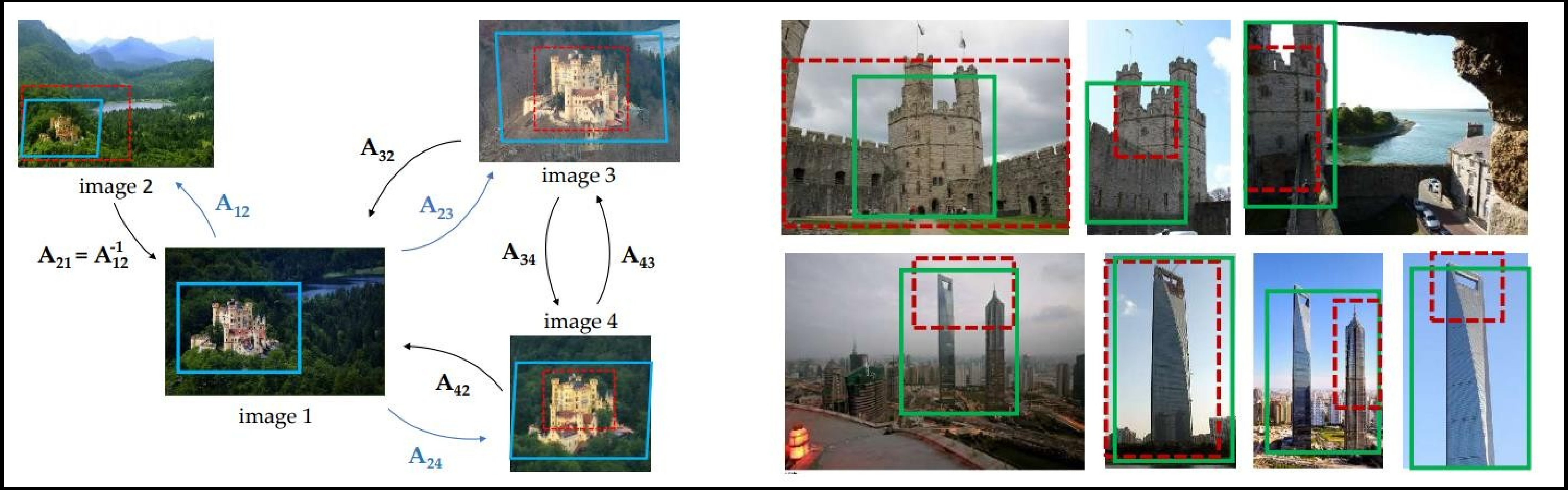
Positive examples: images that share 3D points with the query
Hard positives: positive examples not close enough to the query

anchor | top 1 by CNN | top 1 by inliers | random from top k by inliers

[Radenovic et al. PAMI'19]

Class labels + cleaning

Use classical computer vision to collect training data:
→ Bag-of-Words and spatial verification



[Gordo et al. IJCV'18]

Benchmarks

Instance retrieval (buildings, landmarks)

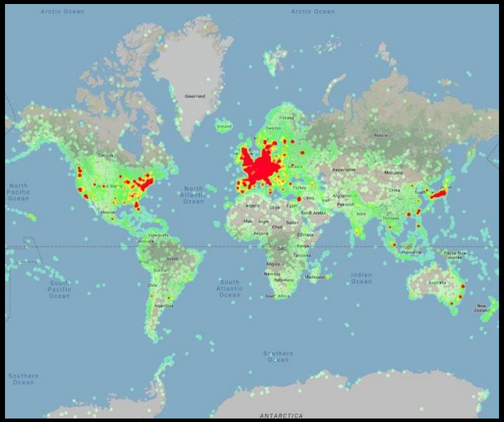
Manually constructed ground truth

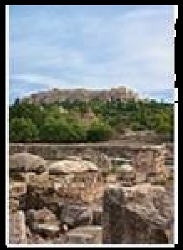
- Oxford buildings [Philbin et al., CVPR’07]
- Paris [Philbin et al., CVPR’08]
- Oxford/Paris revisited + 1M distractors [Radenovic et al., CVPR’18]

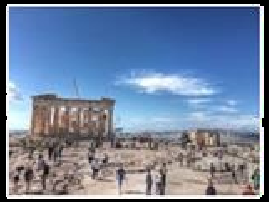
http://cmp.felk.cvut.cz/revisitop/

Landmark recognition and retrieval

Crowd-sourced ground truth







Google Landmarks Dataset

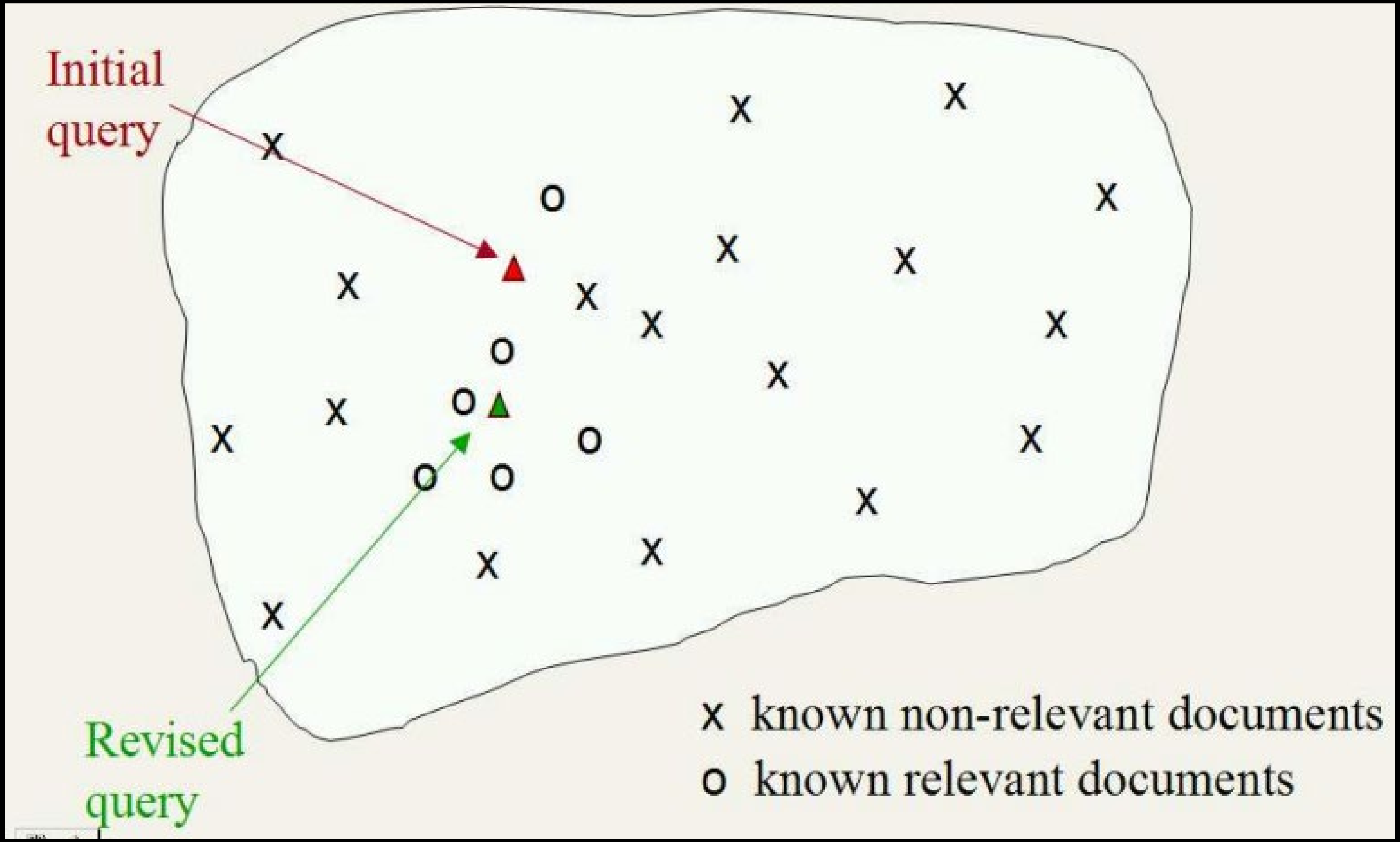
https://github.com/cvdfoundation/google-landmark

- Recognition training set
  4.1m images
  200k landmarks
- Retrieval index set
  762k images (1/3 decrease)
  101k landmarks
- Test set
  118k images
  about 1% depicts landmarks

Post-processing on online time

Query expansion

Use NN information to get more confident query.



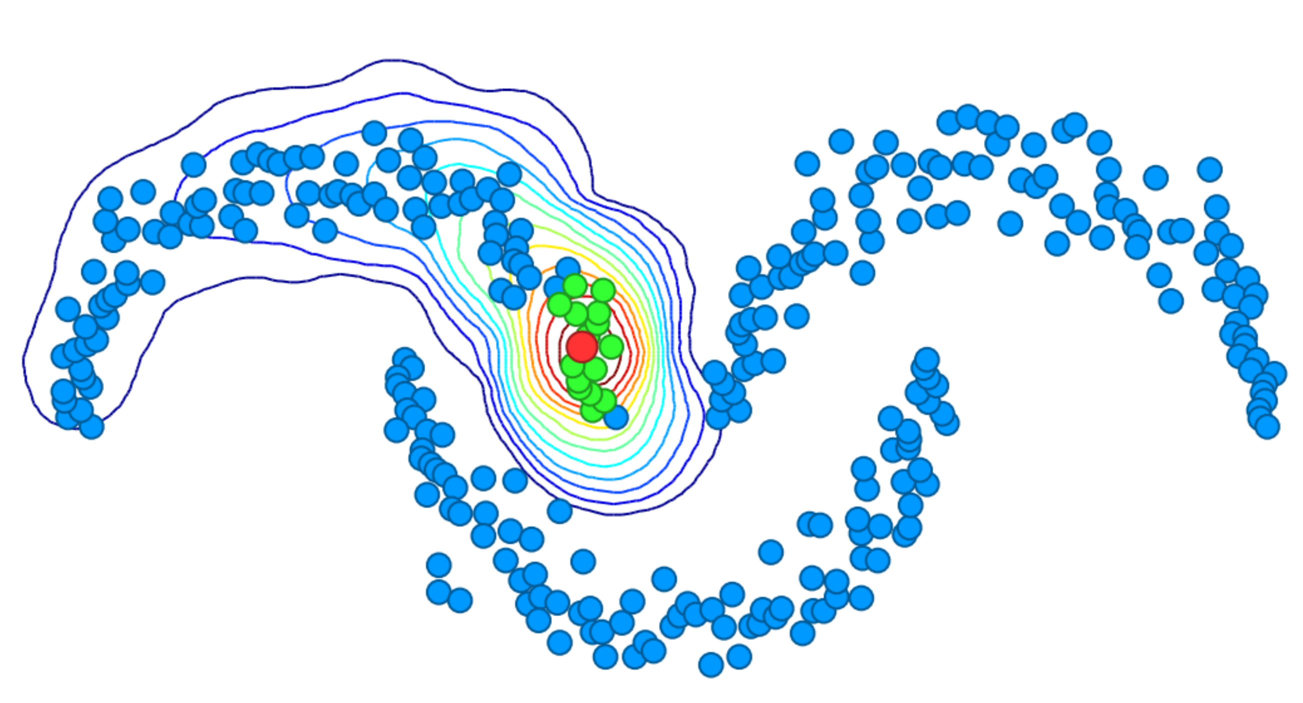
[Chum et al. ICCV'07]

Diffusion(random walk) on feature space

High dimensional feature is likely to have a manifold shape.

$$\mathbf{f}^t = \alpha S \mathbf{f}^{t-1} + (1-\alpha)\mathbf{y}.$$

Iterative manner with affinity graph



[Iscen et al. CVPR'19]

Performance comparison