Deep Learning based Image Search

Sung-eui Yoon
윤성의

Announcements

- There are only 6 students in the class
  - You can do a single-man project or projects w/ 2 people
  - You can bring your own research as long as it is clearly related to the course theme
- Each student
  - Give two talks; each talk time is 15 min
  - Each talk covers one main paper with related papers
- Each team
  - Give a mid-term review presentation for the project
  - Give the final project presentation

Schedule

- Apr-17 (Wed): mid-term exam

- Apr 22, 24, 29, Paper Presentation I
- May 1, 8 Mid-term Project Presentation
- May 13 (no class due to ICRA):

- May 20, 22, 27 Paper Presentation II
- May 29, Reserved
- Jul, 3, 5: Final-term project presentation
- Jul, 10, 12 Reserved (final exam)

Deadlines

- **Declare project team members**
  - **At the class of 3/20 ~~at KLMS~~**

- **Confirm schedules of paper talks and project talks at 3/20**

- **Declare two papers for student presentations**
  - **by 3/26 at KLMS**
  - **Discuss them at the class time of 3/27**

Deadlines

- Declare project team members
  - At the class of 3/20 ~~at KLMS~~
- Confirm schedules of paper talks and project talks at 3/20
- Declare two papers for student presentations
  - by 3/26 at KLMS
  - Discuss them at the class time of 3/27

Class Objectives

- Deep learning based image search
  - CNN based image descriptors
  - Training losses, and data
  - Benchmarks

- At last time, we discussed:
  - Scale invariant region selection
  - SIFT as a local descriptor

Learning-based methods

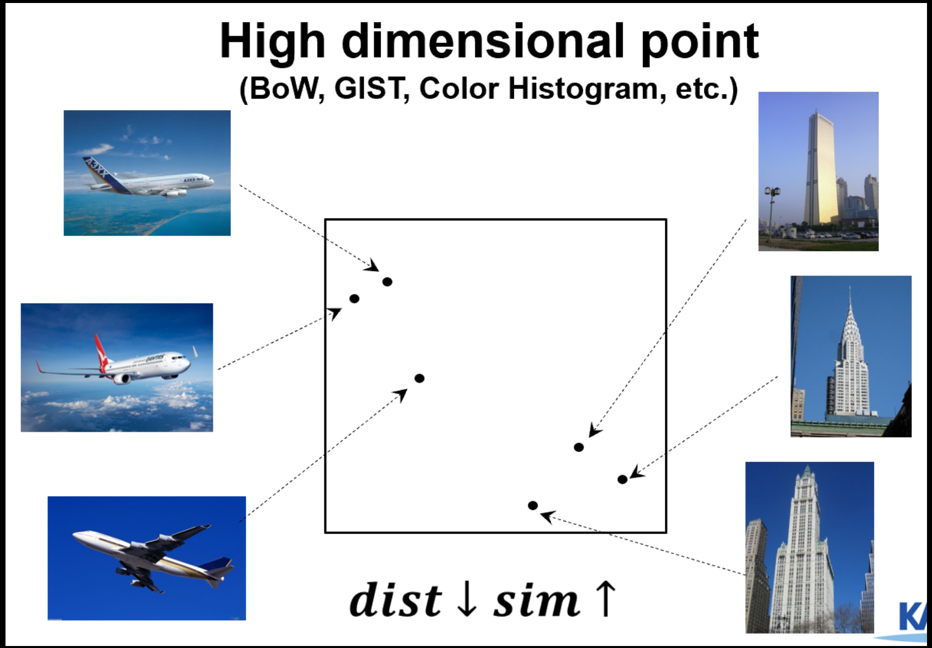
Most of this presentation materials was built upon Tolias's, and prepared by TA

Ack.: Jaeyoon Kim (김재윤)

Global descriptor

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

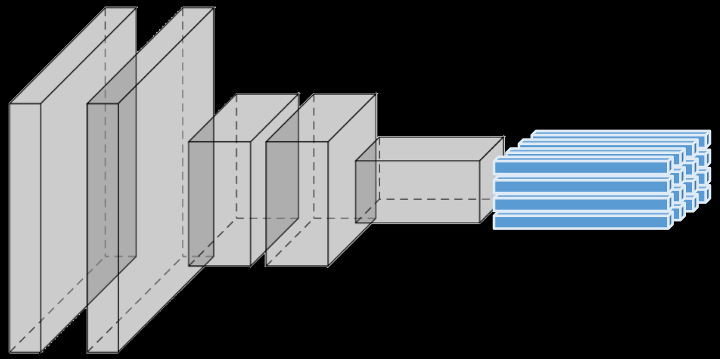
global desc. *embedding function (e.g., Neural Network)*



- Instance search reduces to similarity search in d-dimensional space

- Compatible with efficient nearest neighbor techniques

Global descriptors with CNNs



fully convolutional network

Embedding & aggregation

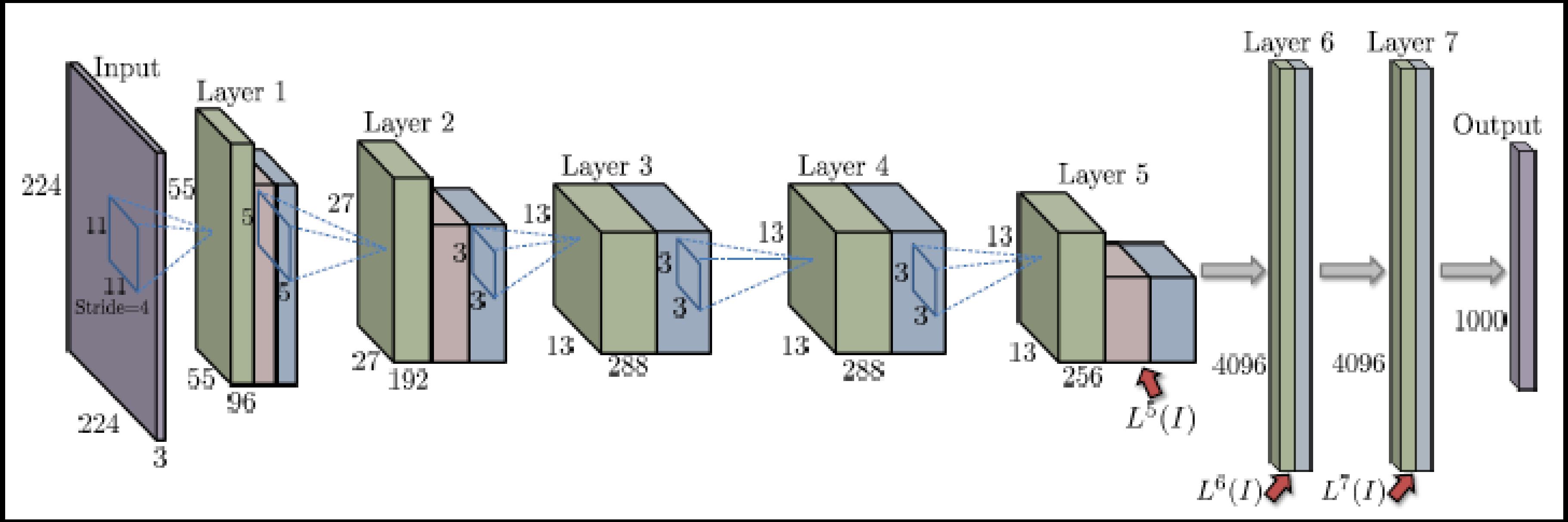
Global desc. by aggregation g():

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

$\mathcal{X}$: a set of local descriptors

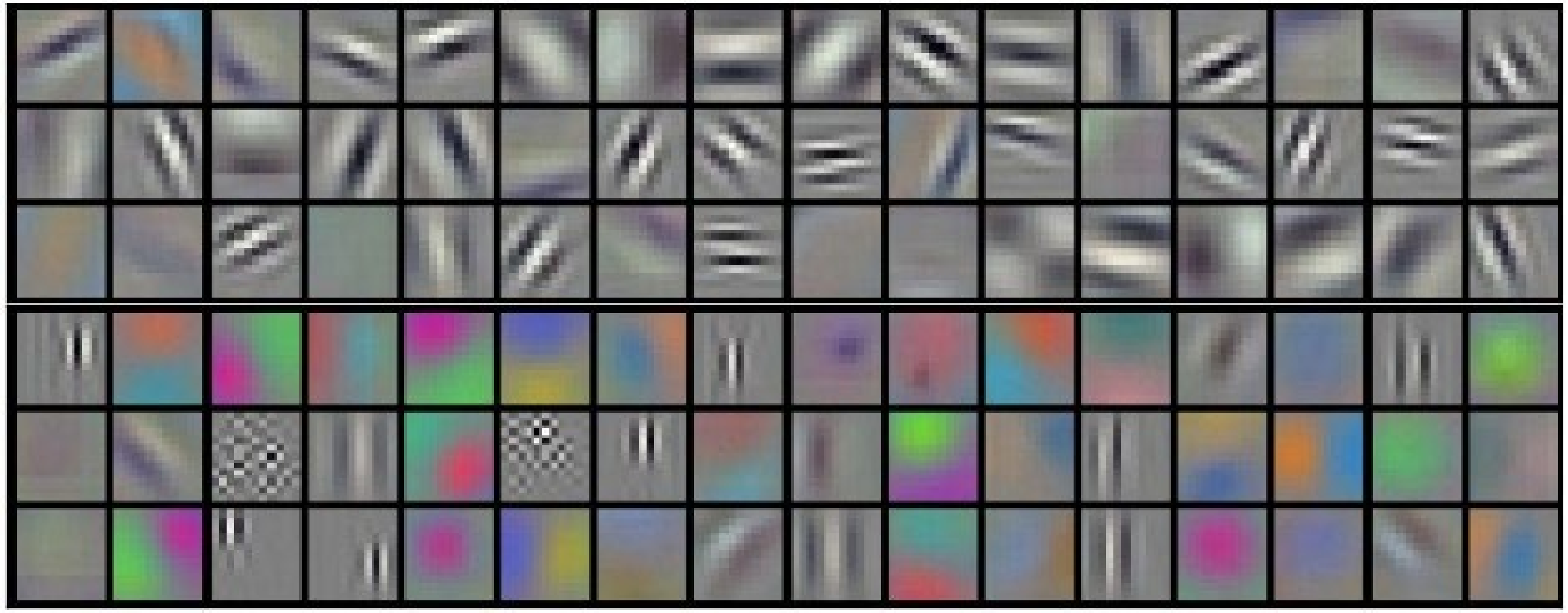
Neural Codes for Image Retrieval [ECCV 14]

Uses top layers of CNNs as high-level global descriptors (Neural Codes) for image search



Visualizing filters

- Example: filter in the first layer of AlexNet

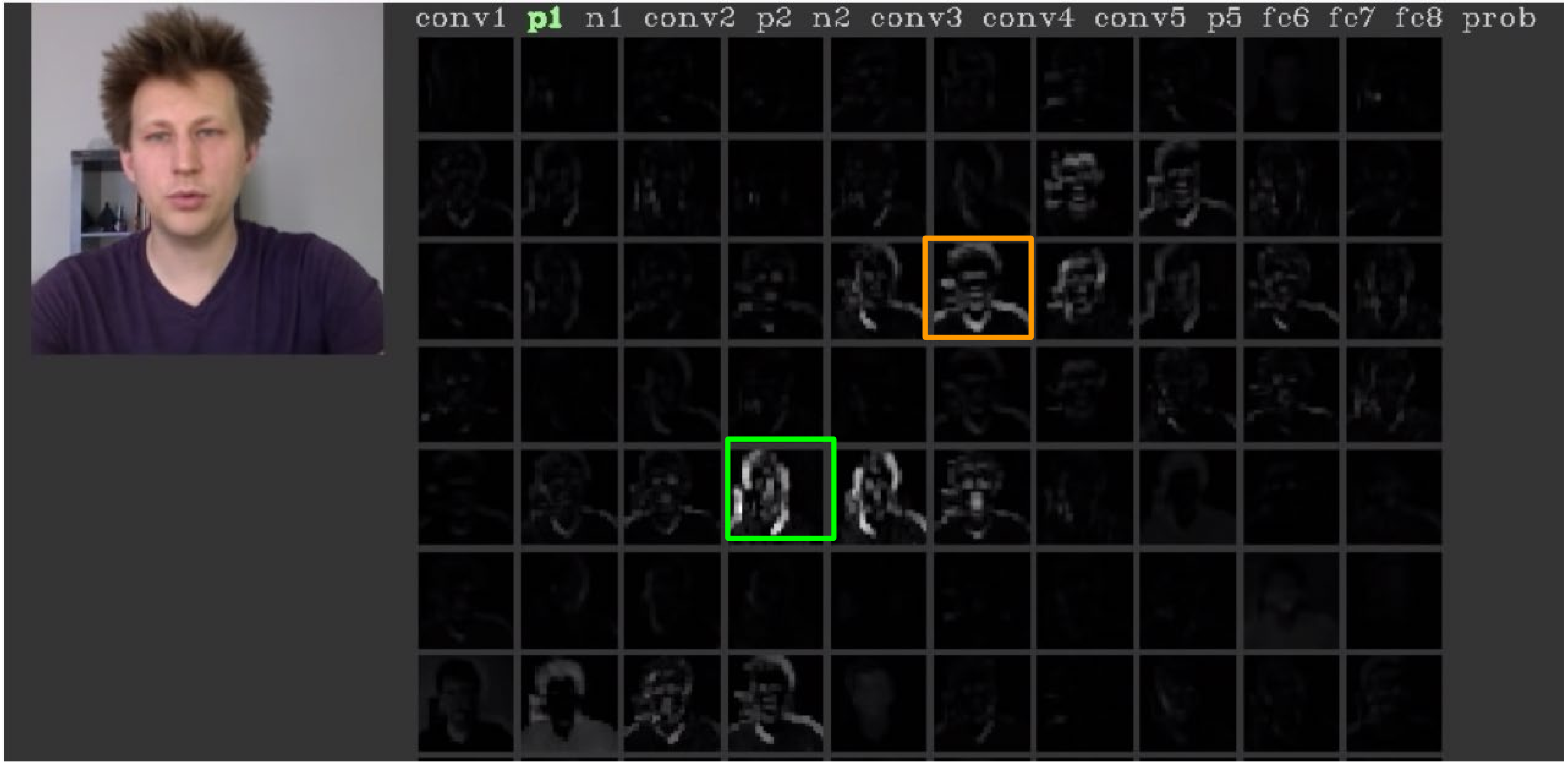


Edges in various angle
(horizontal, vertical, diagonal, etc.)

Color patterns
(green, magenta, etc.)

Visualizing activation map

- **Conv1** feature map in AlexNet



Strong activation around object boundary (edge)

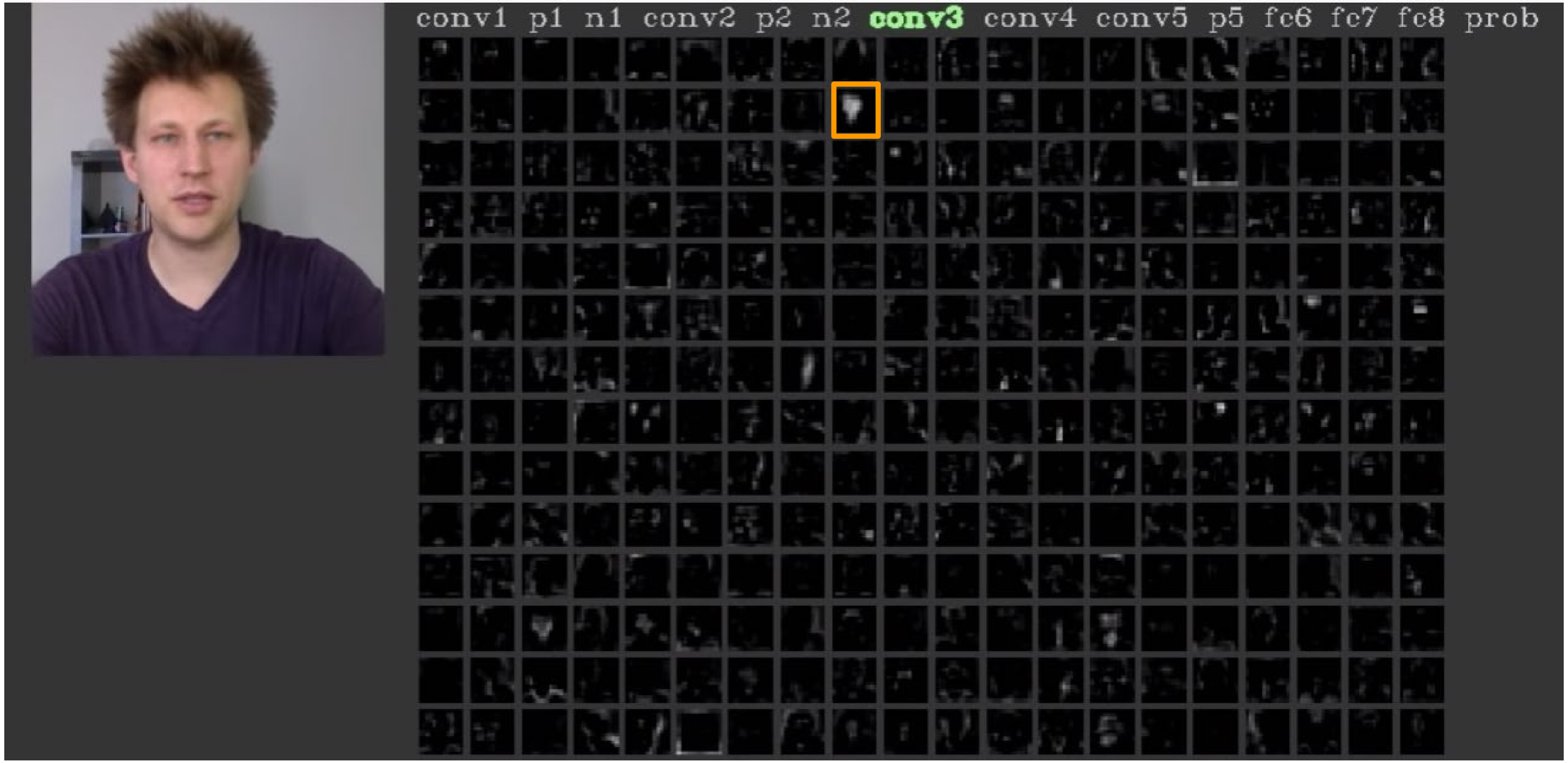
E.g. horizontal

E.g. vertical

Yosinki et al., Understanding Neural Networks Through Deep Visualization

Visualizing activation map

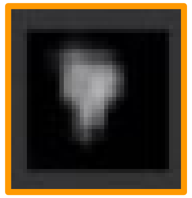
- **Conv3** feature map in AlexNet



Yosinki et al., Understanding Neural Networks Through Deep Visualization

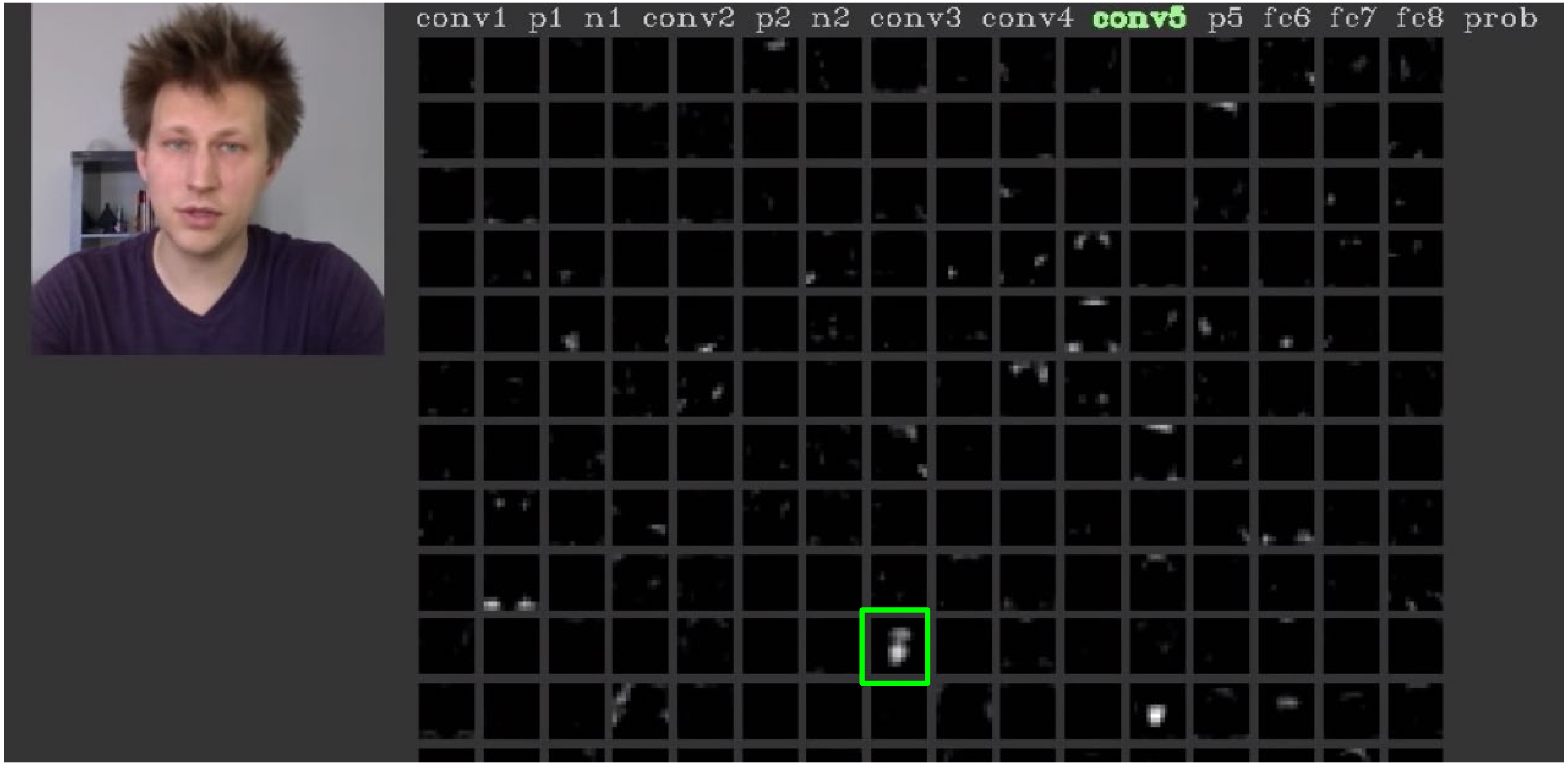
Strong activation in more meaningful groups

E.g. Skin colors



Visualizing activation map

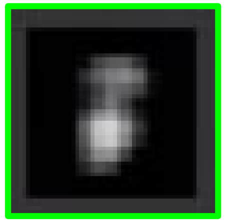
- **Conv5** feature map in AlexNet



Yosinki et al., Understanding Neural Networks Through Deep Visualization

Strong activation in more meaningful groups

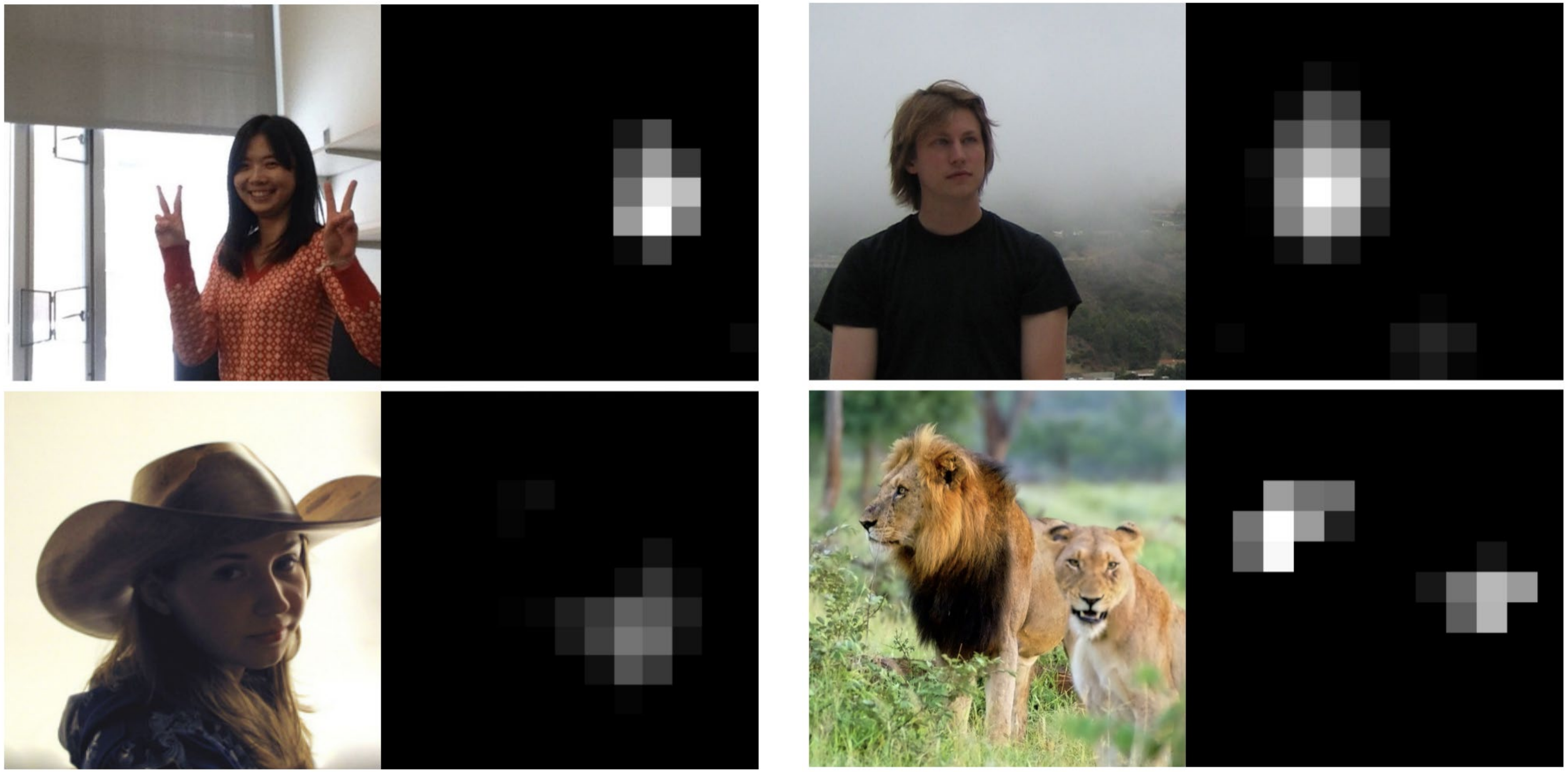
E.g. Face



It's also much sparse → activated for more larger semantic groups

Visualizing activation map

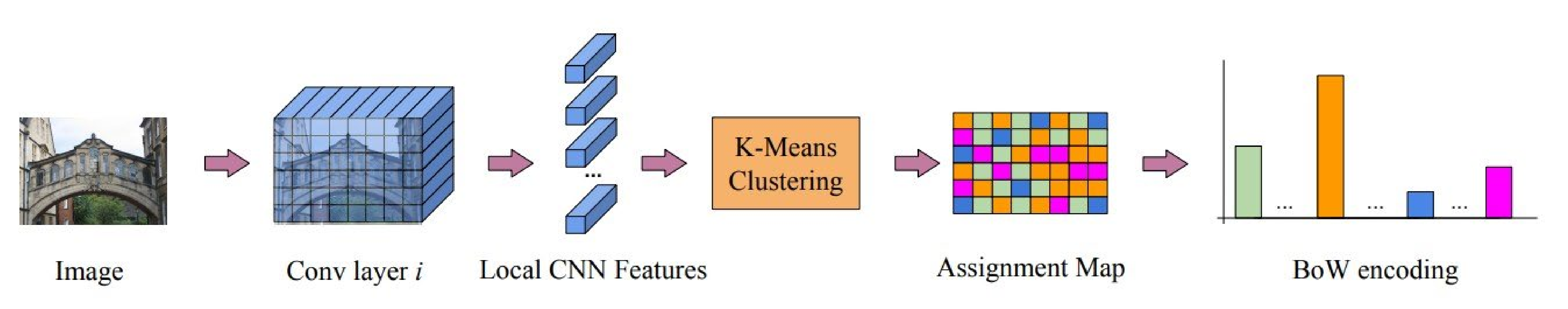
- 151th filter at conv5 layer does the face detection!



Yosinki et al., Understanding Neural Networks Through Deep Visualization

BoW (Bag-of-visual-Words) with CNN features

- Inspired by a classical BoW approach; a type of aggregation
- Less commonly used now



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

[Mohedano et al. ICMR'16]

Sum pooling – SPoC (Sum Pooling w/ Center prior) descriptor

- Descriptor by simple summation

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

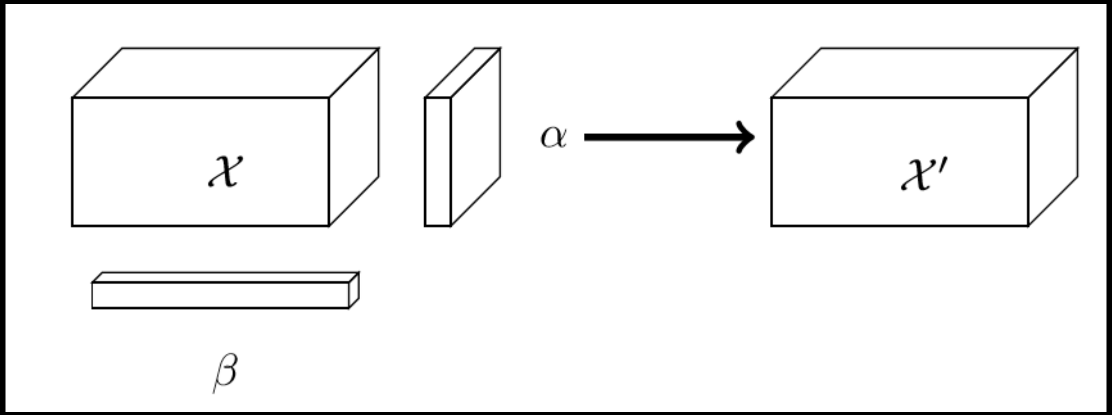
- Pair-wise similarity of two images; dot product, cosign similarity

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

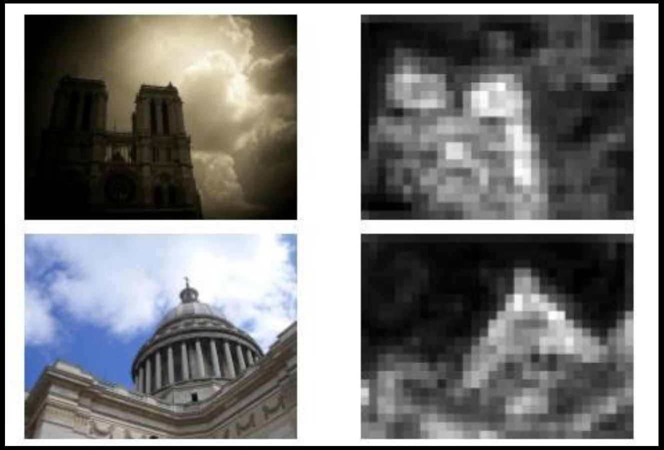
- Simple but works
  → discriminative power of CNN activations

[Babenko & Lempitsky, ICCV’15]

Weighted sum pooling – CroW (Cross-dim. Weighted) descriptor



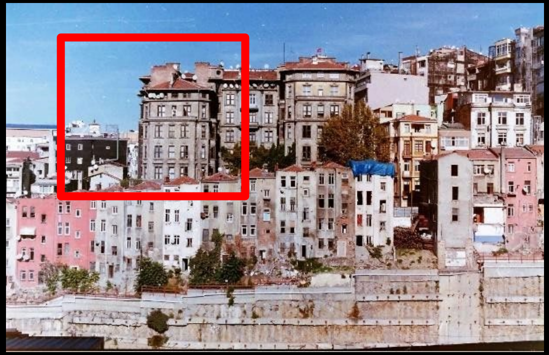
α: weight based on L2 norm of local descriptors
β: channel-wise attention



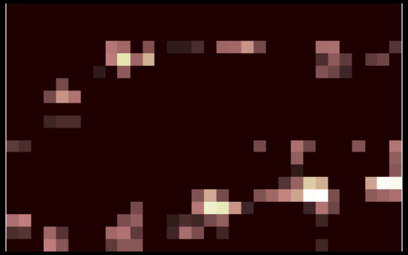
example of α

[Kalantidis et al., ECCV'16]

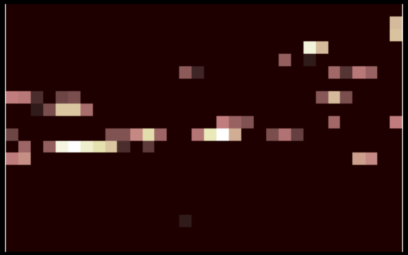
Max pooling – MAC (Max. Activation of Conv.) descriptor



Input image

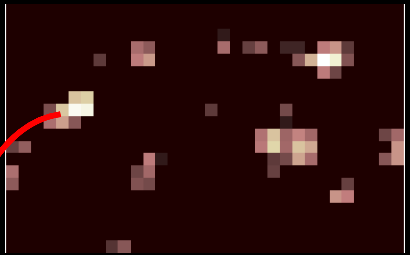


conv_5 filter 1



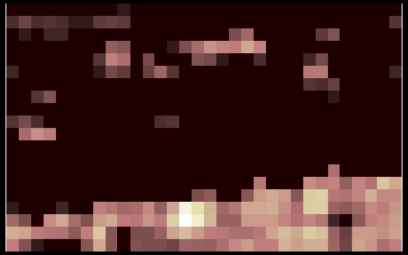
conv_5 filter 2

….



conv_5 filter i

….



conv_5 filter K

maximum activation

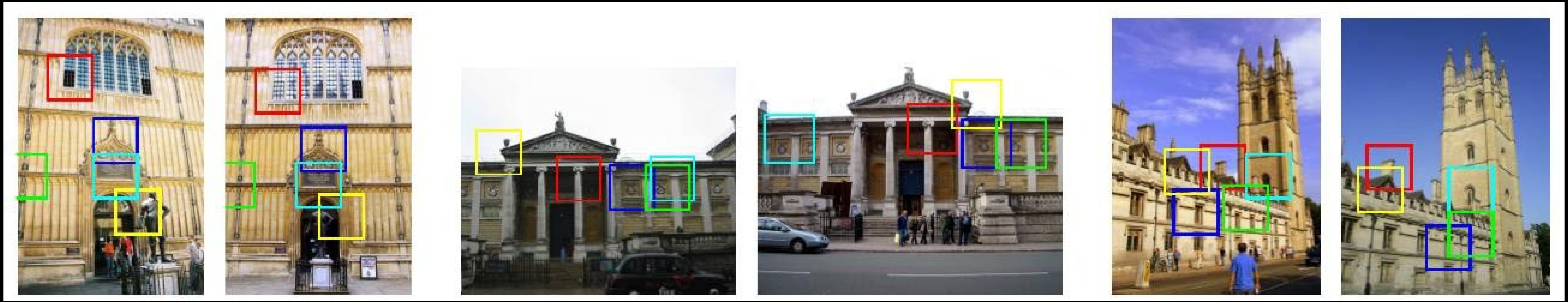
$$\mathrm{MAC} = [f_1, \ldots, f_i, \ldots, f_K]$$

f_i: a scalar value of corresponding position at filter i

[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

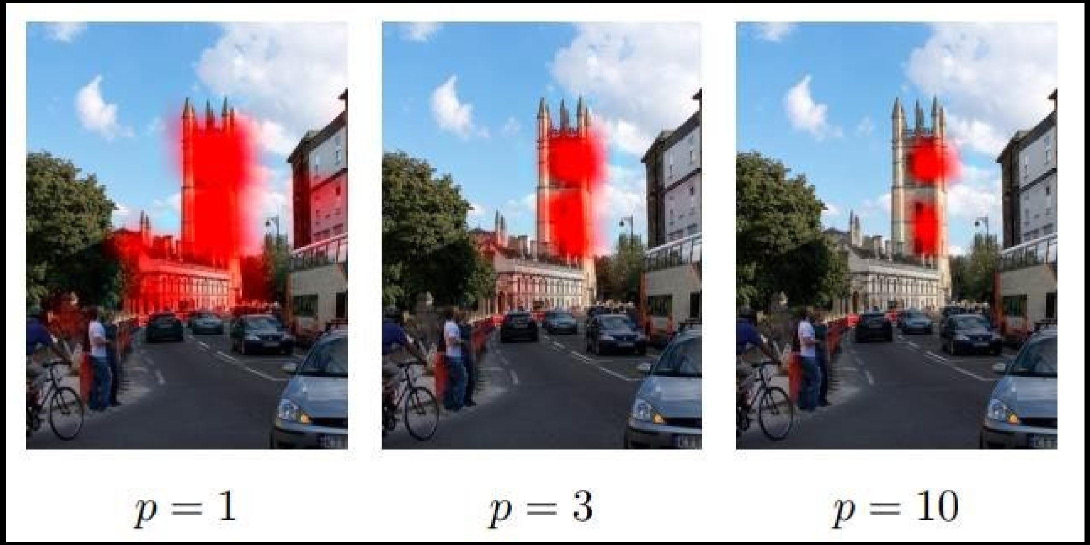
[Razavian et al., MTA'16] [Tolias et al., ICLR'16]

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

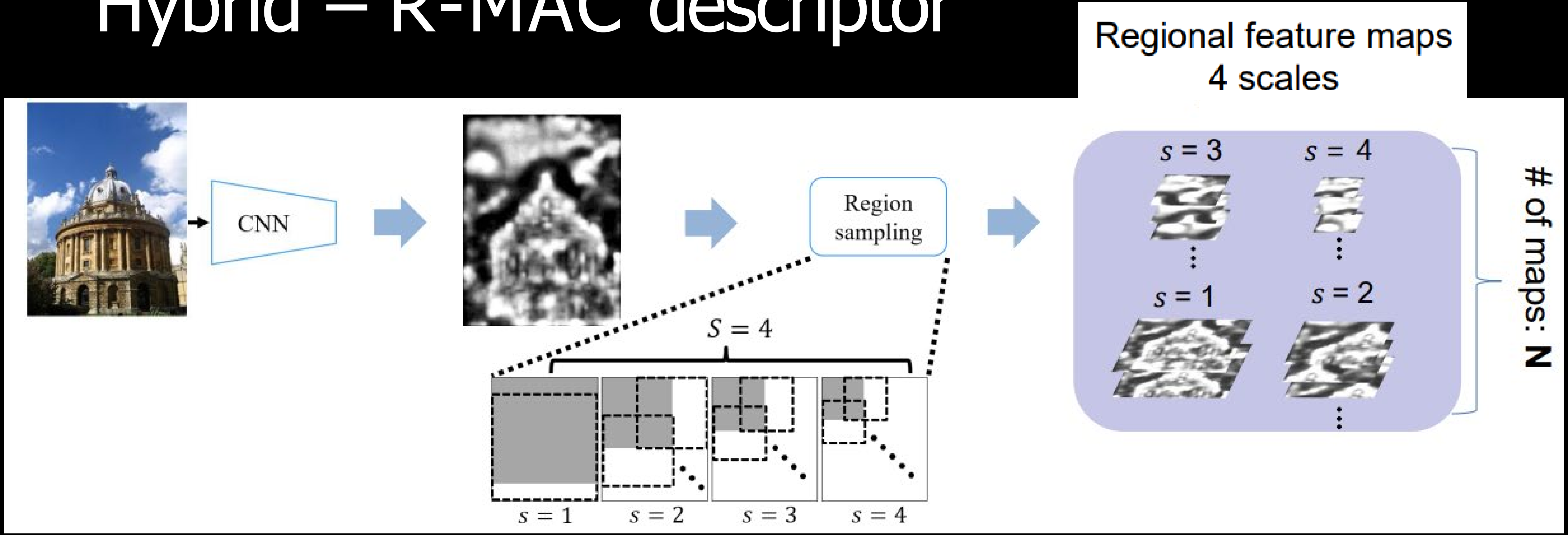
$p \rightarrow \infty$ max pool (MAC)
$p = 1$ avg pool (SPoC)



Red regions show activated regions

[Radenovic et al., PAMI’19]

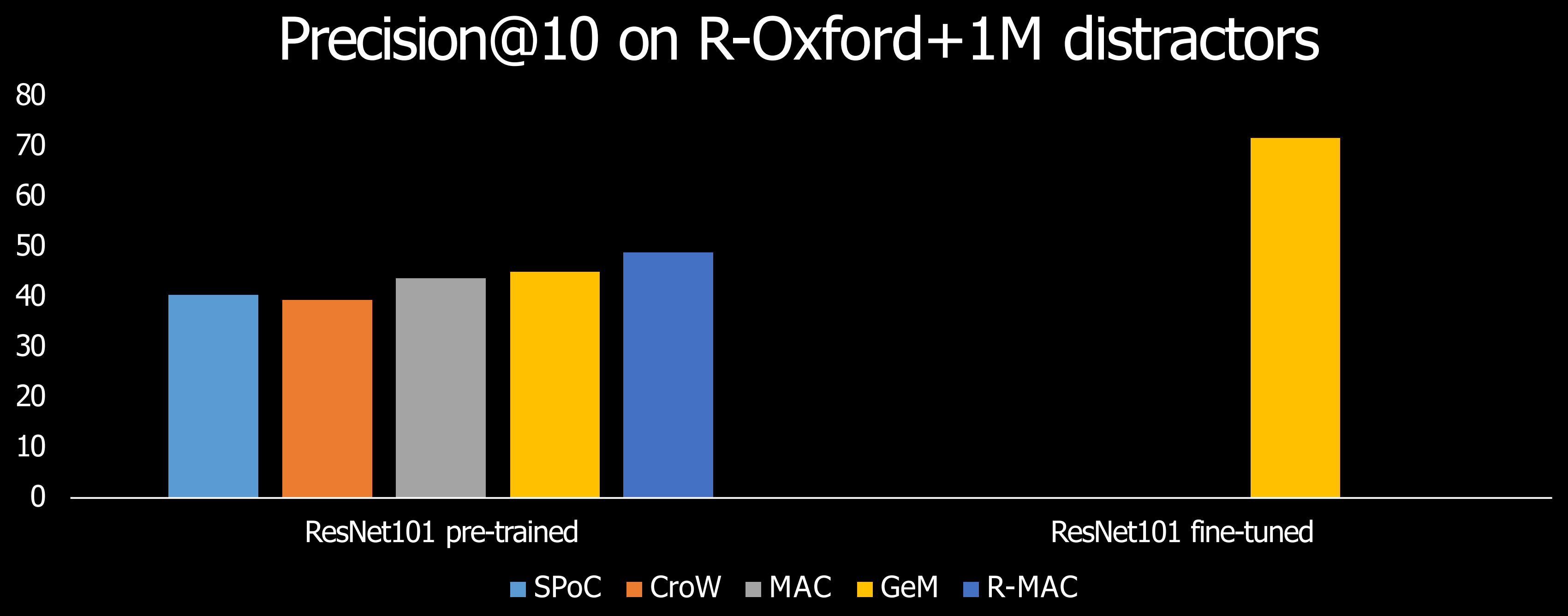
Hybrid – R-MAC descriptor



- Multi-scale region sampling
- Sum aggregate over regions

[Tolias et al., ICLR’16]

Performance comparison



Fine-tuning improvement for GeM: +26.6%

Fine-Tuning for Search

- **Use CNN features that were trained with ImageNet**
- **Retraining with a task-specific dataset achieve higher accuracy**
  - **Can lower accuracy when using dissimilar datasets**

Fine-Tuning for Search

Results before & after retraining

Neural codes trained on ILSVRC					
Layer 5	9216	0.389	—	0.690*	3.09
Layer 6	4096	0.435	0.392	0.749*	3.43
Layer 7	4096	0.430	—	0.736*	3.39
After retraining on the Landmarks dataset					
Layer 5	9216	0.387	—	0.674*	2.99
Layer 6	4096	0.545	0.512	**0.793***	3.29
Layer 7	4096	0.538	—	0.764*	3.19
After retraining on turntable views (Multi-view RGB-D)					
Layer 5	9216	0.348	—	0.682*	3.13
Layer 6	4096	0.393	0.351	0.754*	3.56
Layer 7	4096	0.362	—	0.730*	3.53

Ack.: Neural Codes for Image Retrieval

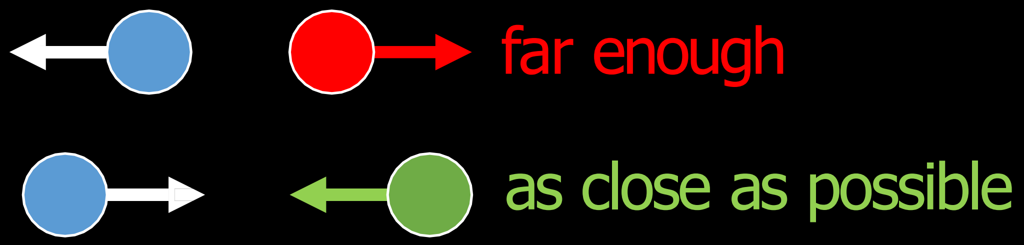
Landmark dataset has similar images to Oxford

Training loss

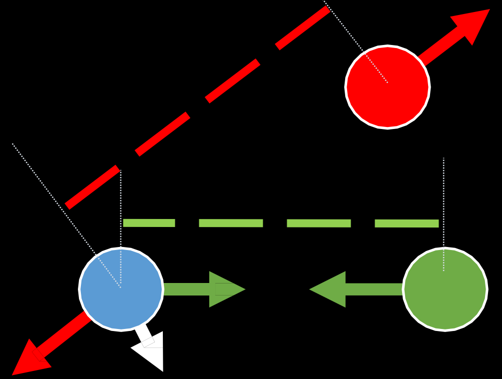
Loss functions for metric learning

anchor negative positive

Contrastive loss



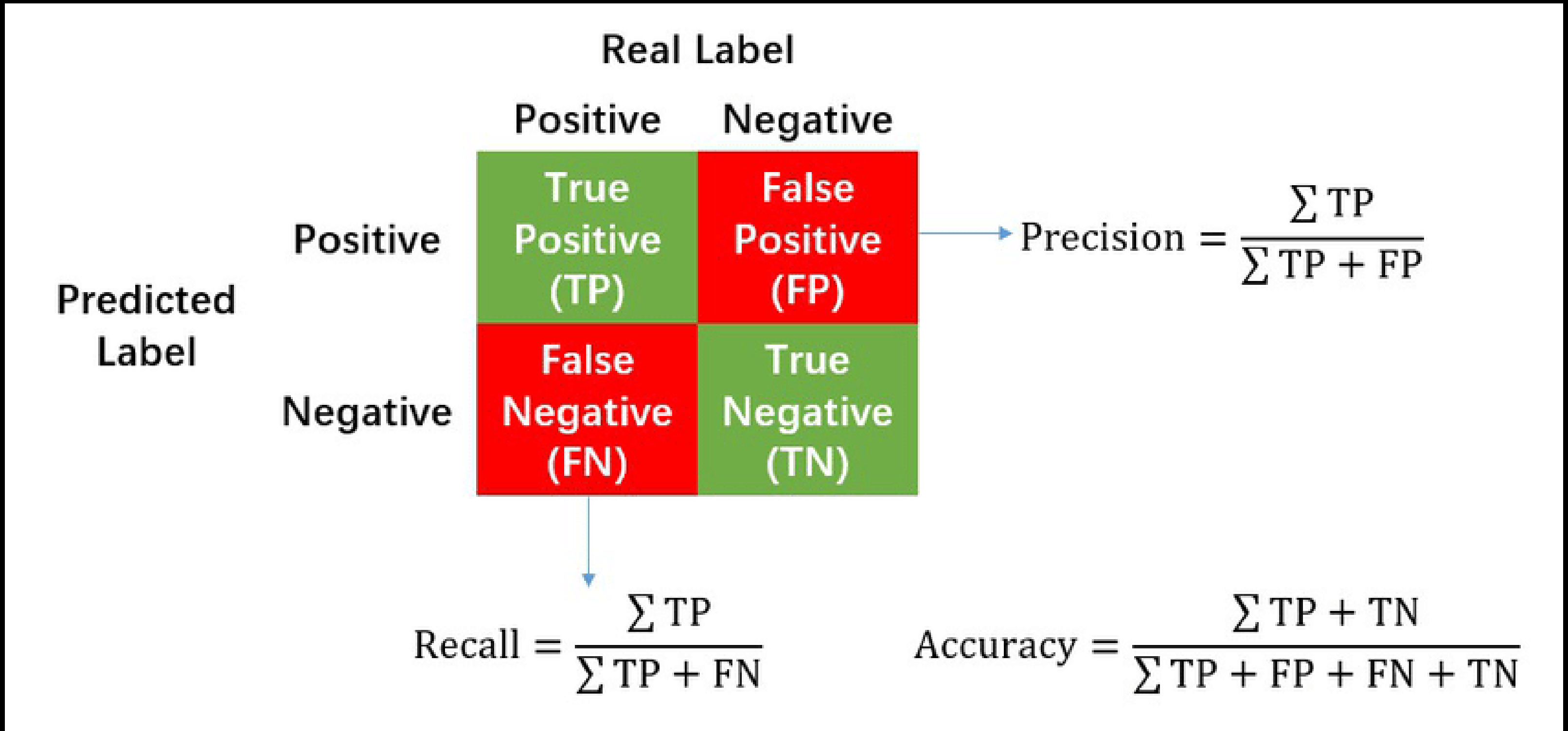
Triplet loss



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

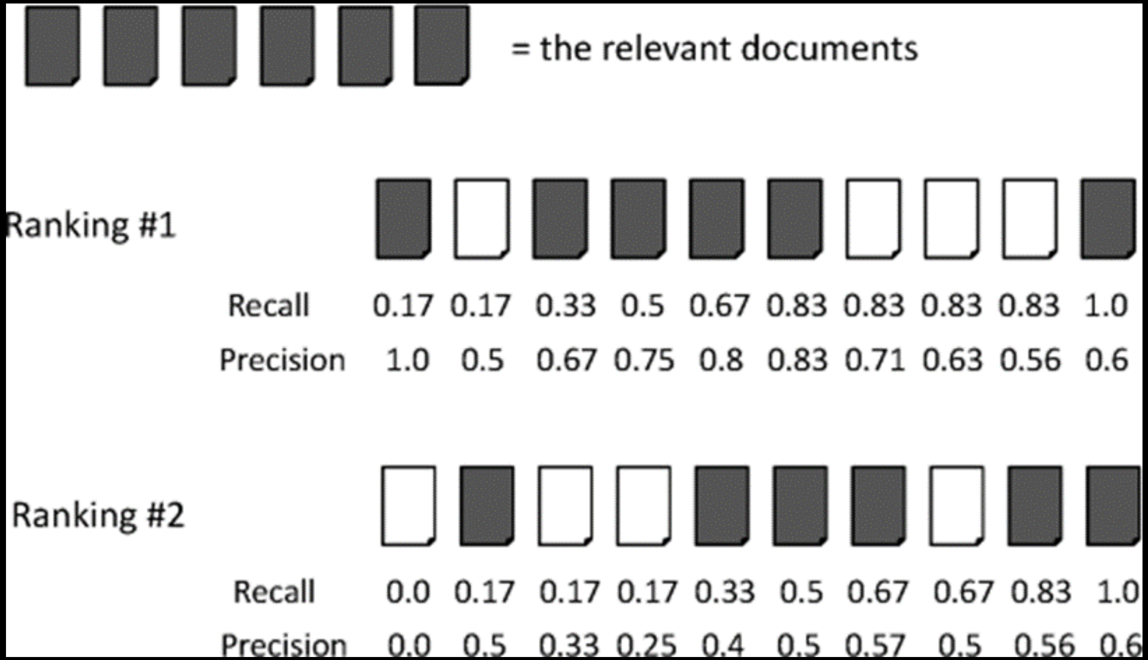
Average Precision loss

- Definition of recall and precision



Average Precision loss

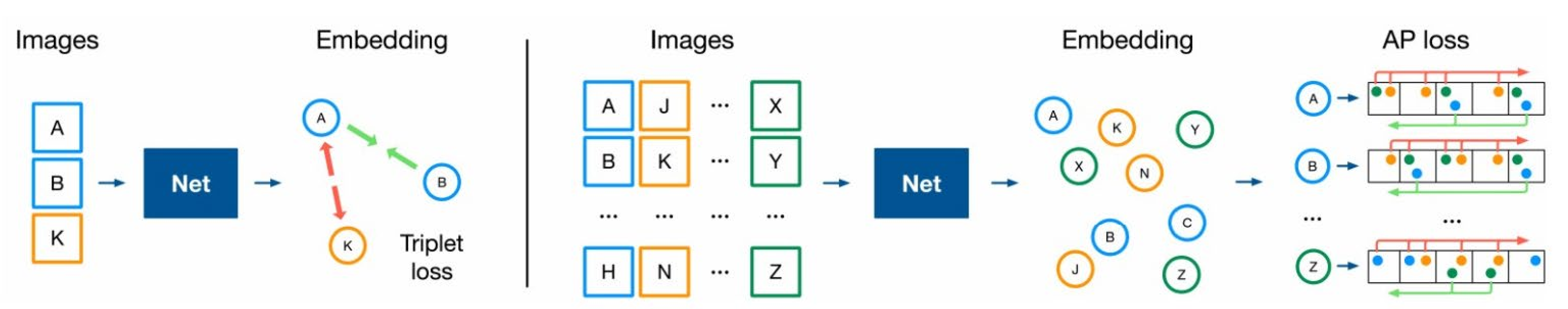
- Two examples of average precision



$$\text{Ranking \#1: } (1.0 + 0.67 + 0.75 + 0.8 + 0.83 + 0.6)/6 = 0.78$$

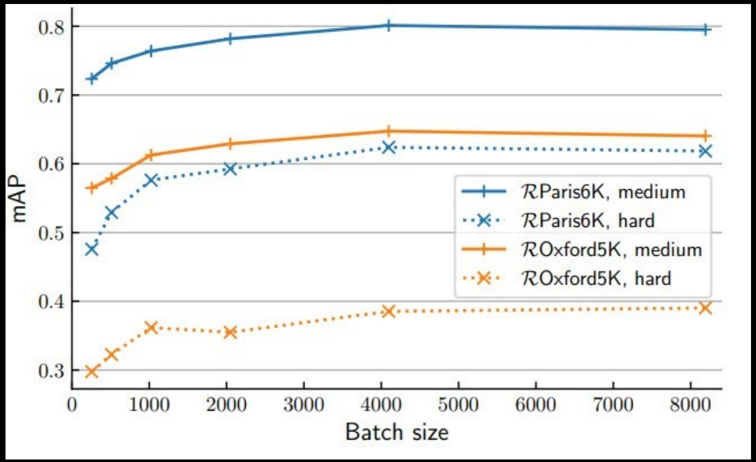
$$\text{Ranking \#2: } (0.5 + 0.4 + 0.5 + 0.57 + 0.56 + 0.6)/6 = 0.52$$

Average Precision loss



Same colors indicate positive pairs

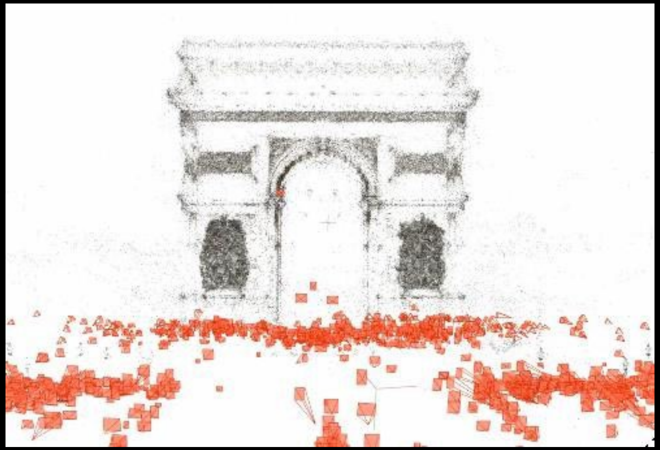
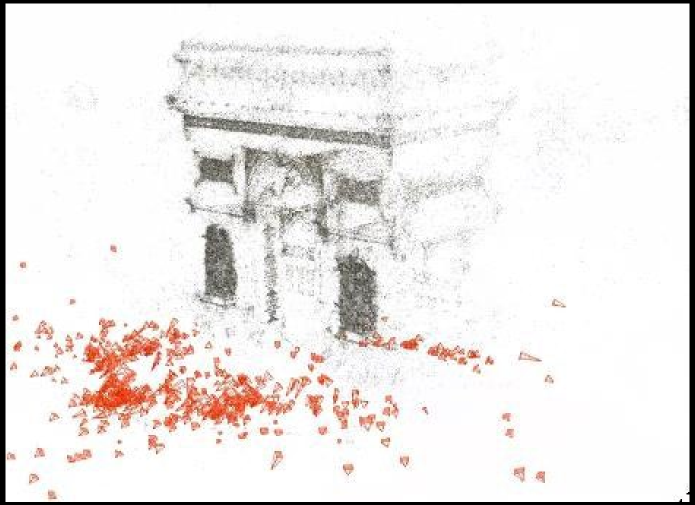
The larger the batch the better and less dependency on sampling



 [Revaud et al., ICCV'19]

Training data

Training data from SfM

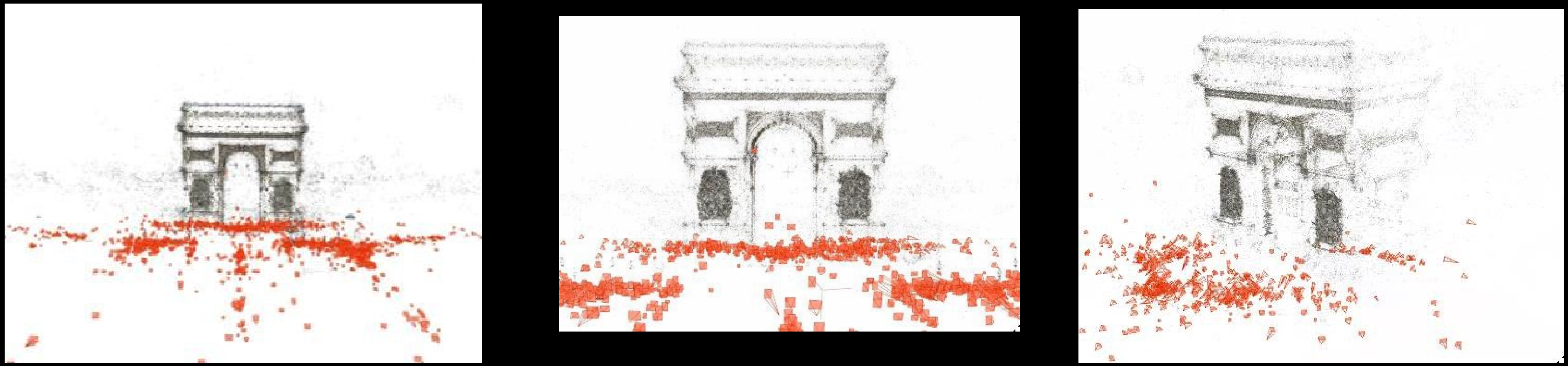




7.4M images → 713 training 3D models

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

Training data from SfM

camera orientation known
number of inliers known

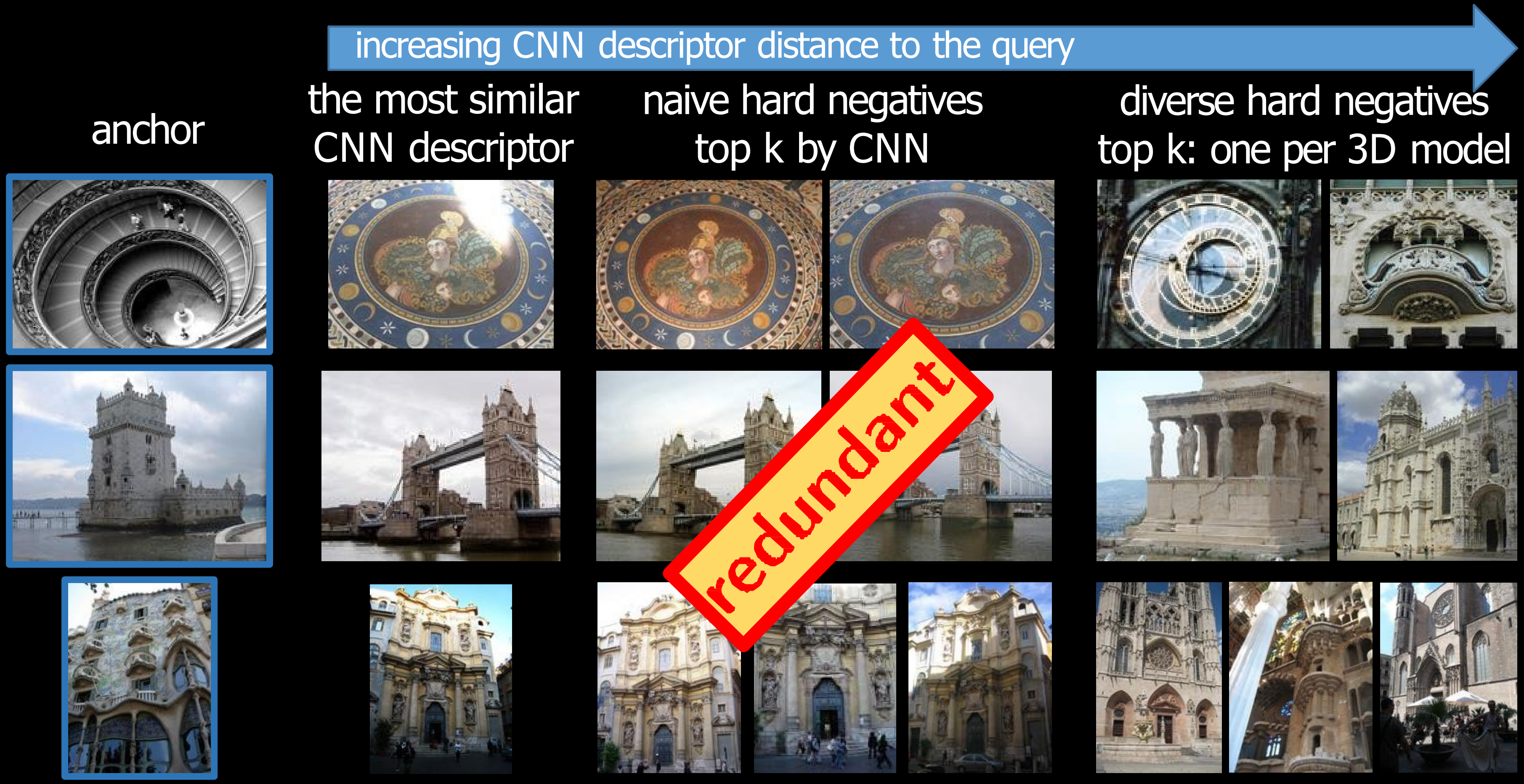


7.4M images → 713 training 3D models

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

Training data from SfM: hard negatives

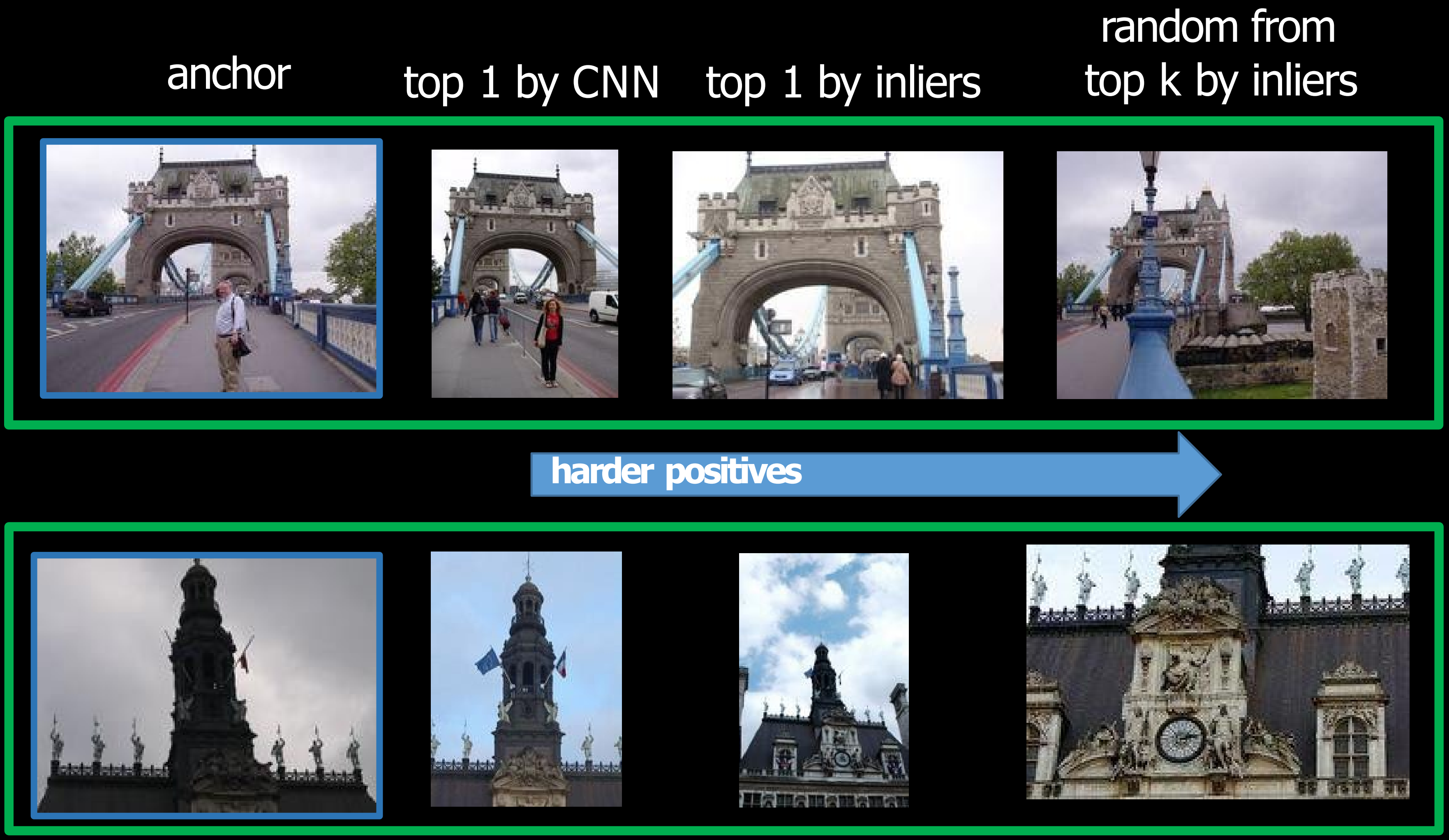
Negative examples: images from different 3D models than the query
Hard negatives: closest negative examples to the query



[Radenovic et al.
PAMI’19]

Training data from SfM: hard positives

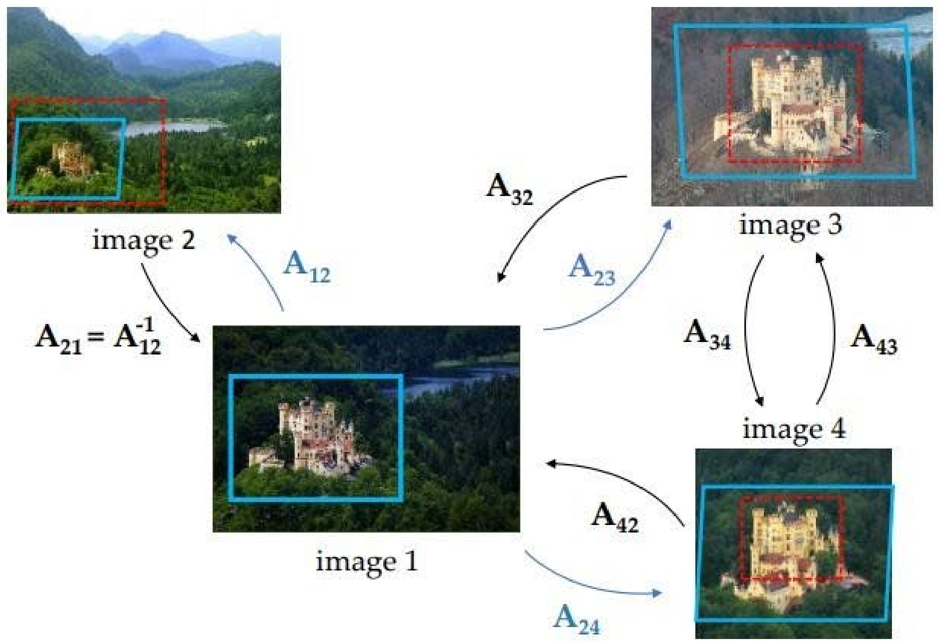
Positive examples: images that share 3D points with the query
Hard positives: positive examples not close enough to the query

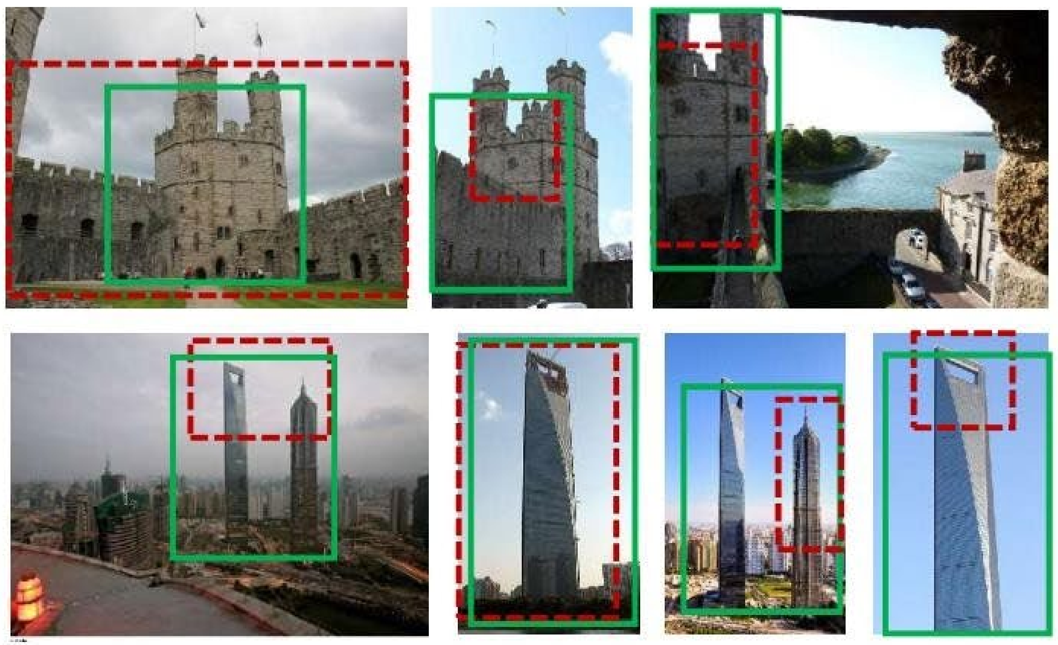


[Radenovic et al. PAMI'19]

Class labels + cleaning

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





A represents affine transformation matrix

[Gordo et al. IJCV'18]

Benchmarks

Instance retrieval (buildings, landmarks)

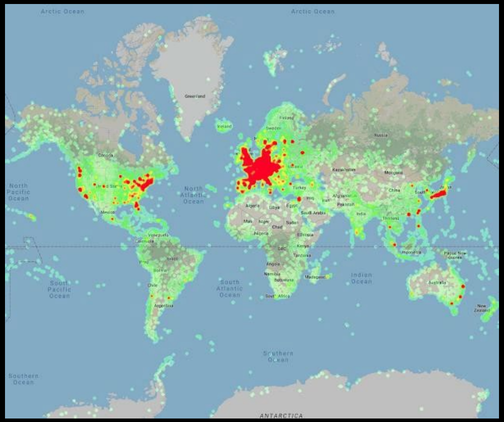
Manually constructed ground truth

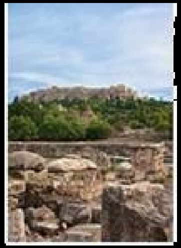
- Oxford buildings [Philbin et al., CVPR'07]
- Paris [Philbin et al., CVPR'08]
- Oxford/Paris revisited + 1M distractors [Radenovic et al., CVPR'18]

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

Landmark recognition and retrieval

Crowd-sourced ground truth





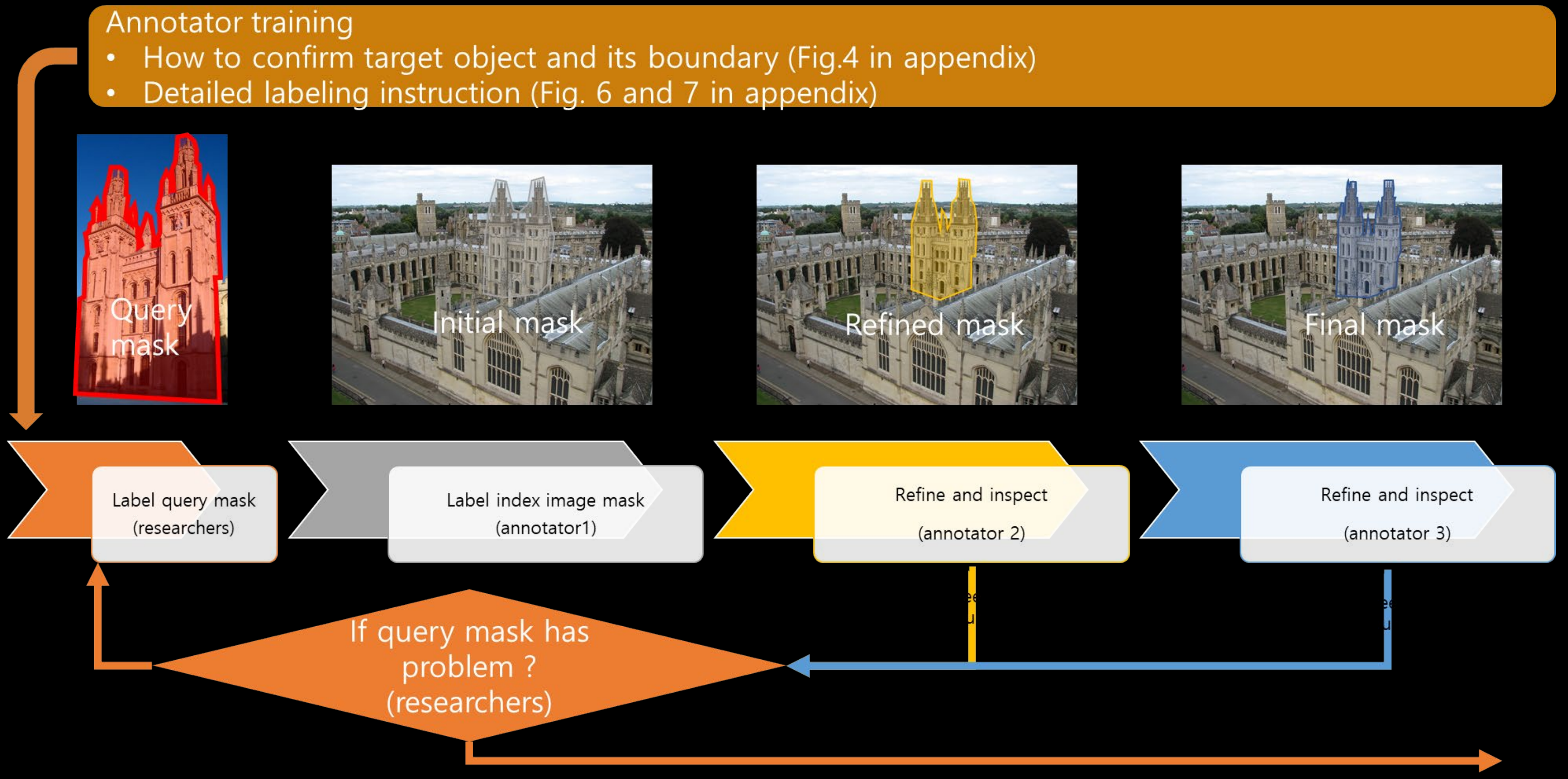
Google Landmarks Dataset

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

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

Pixel retrieval [ICCV 23]

- Benchmark: PROxford/PRParis
  - Use same query and database images with Oxford/Paris
  - Provide pixel-level annotation
- Search pixels that depict the query object from the database



Labeling process

Why pixel retrieval?

Image retrieval

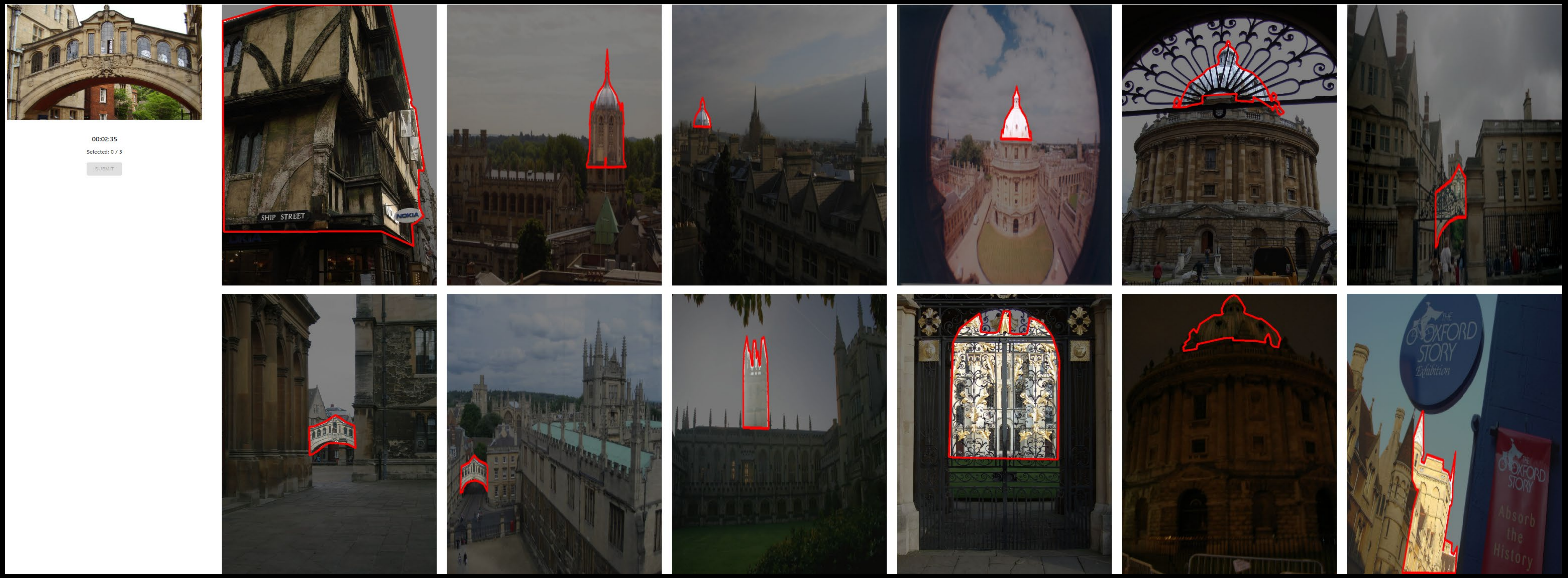
- Search the **images** which contain the query object from the database
- A real-world image has several different objects with complex background
- Users may be difficult to identify the query object from the ranking list

Pixel retrieval

- Search **pixels** that depict the query object from the database
- Retrieve, localize, and segment the target object from the database images

Difficult to check which image is correct.

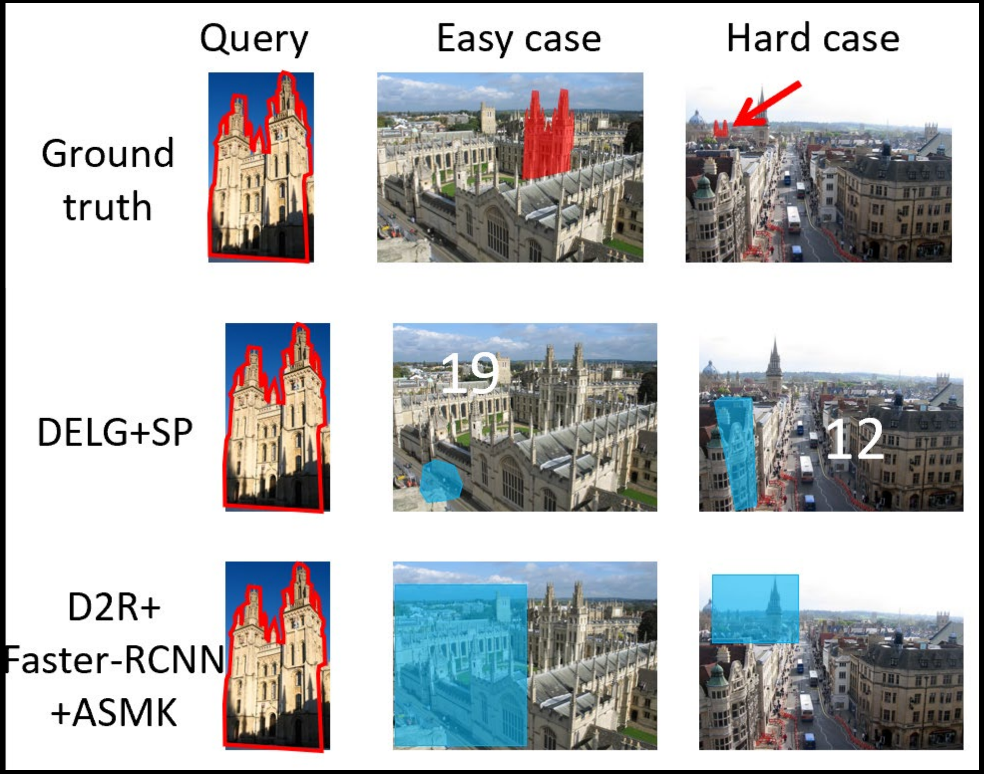
It is easier for users to find the target object if the search engine gives the pixel-level result.

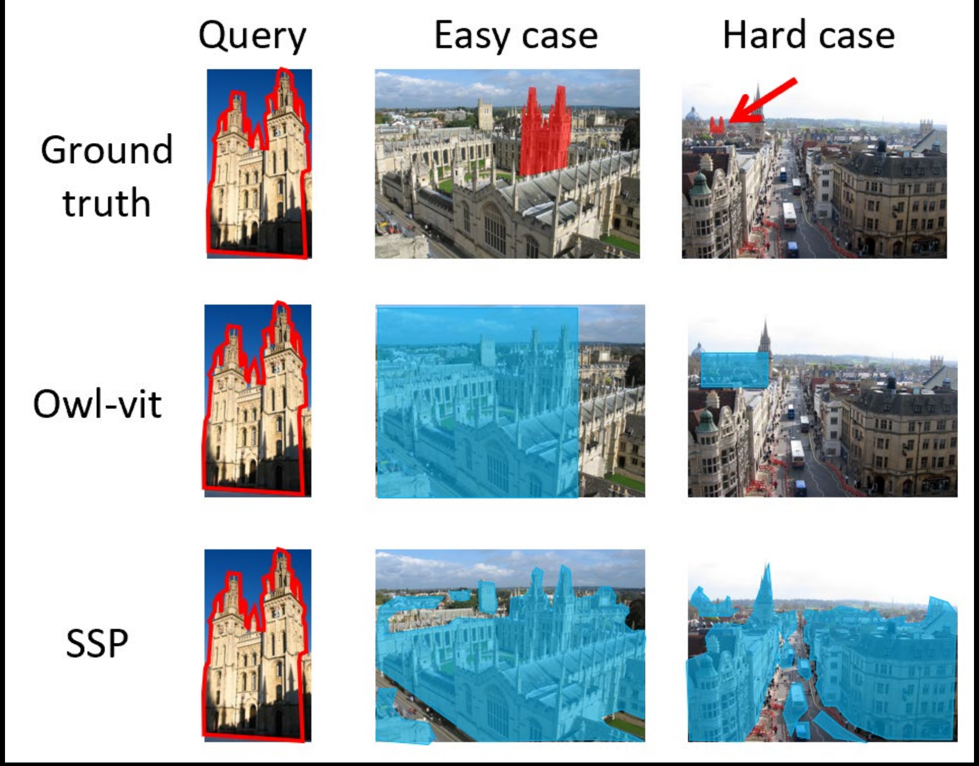


Try more examples: user study

Current SOTA

- Performance of current SOTA methods are not good.
- Need more future studies.





Pixel retrieval performance

PA1

- Understand and implement a basic image retrieval system
- Use a simple UKBenchmark
- Measure its accuracy

Query	First	Second	Third
			

Class Objectives were:

- Deep learning based image search
  - CNN based image descriptors
  - Training losses, and data
  - Benchmarks

Next Time…

- Some post-processing methods and indexing structures

Homework for Every Class

- Go over the next lecture slides
- Come up with one question on what we have discussed today
  - 1 for typical questions (that were answered in the class)
  - 2 for questions with thoughts or that surprised me
- Write questions 3 times before the mid-term exam
  - Write a question about one out of every four classes
  - Multiple questions in one time will be counted as one time
- Common questions are compiled at the Q&A file
  - Some of questions will be discussed in the class
- If you want to know the answer of your question, ask me or TA on person