Visual Search

Hanoi University of
Science and Technology

Diane Larlus

Principal Research Scientist

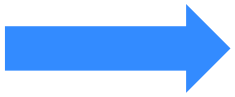
May 31st 2021

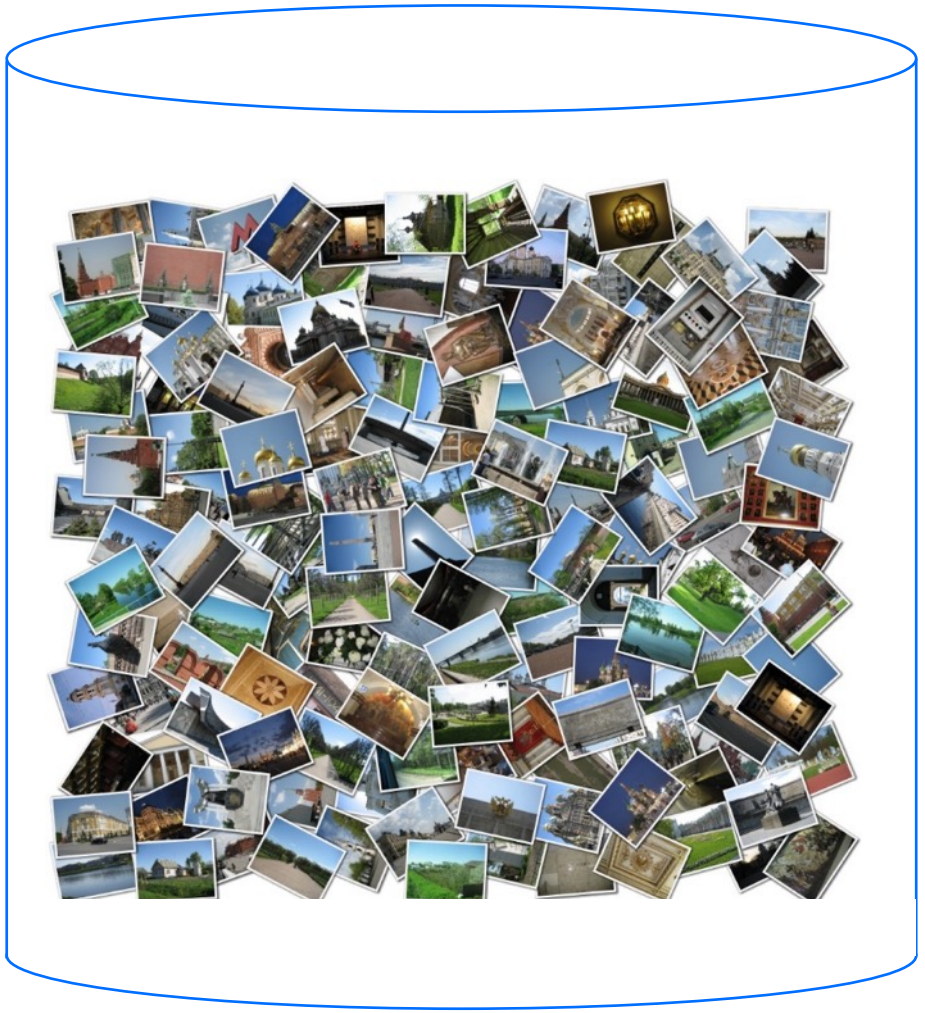
NAVER LABS
Europe

Visual Search

NAVER LABS
Europe

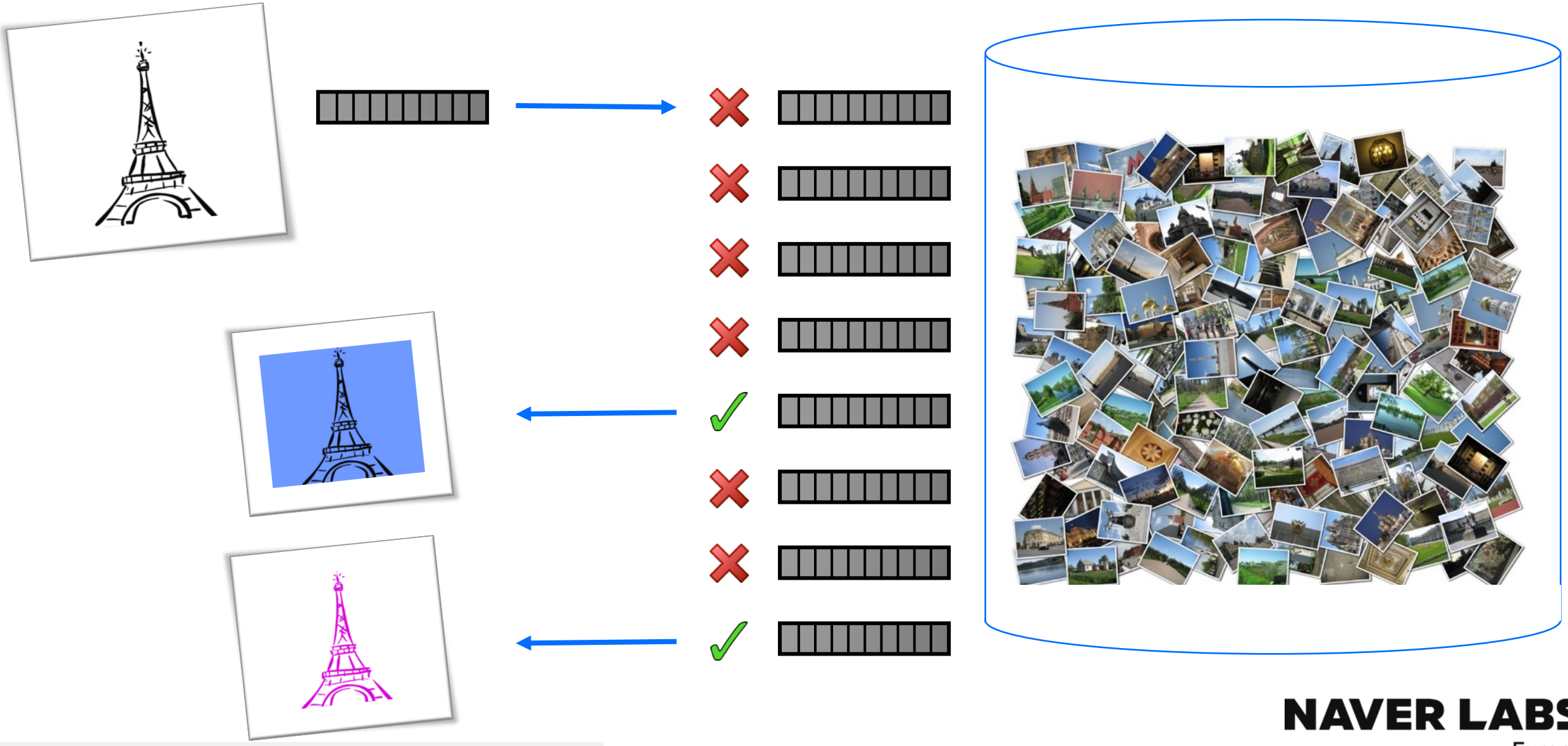
Visual Search - Principle





NAVER LABS Europe

Visual Search - Principle

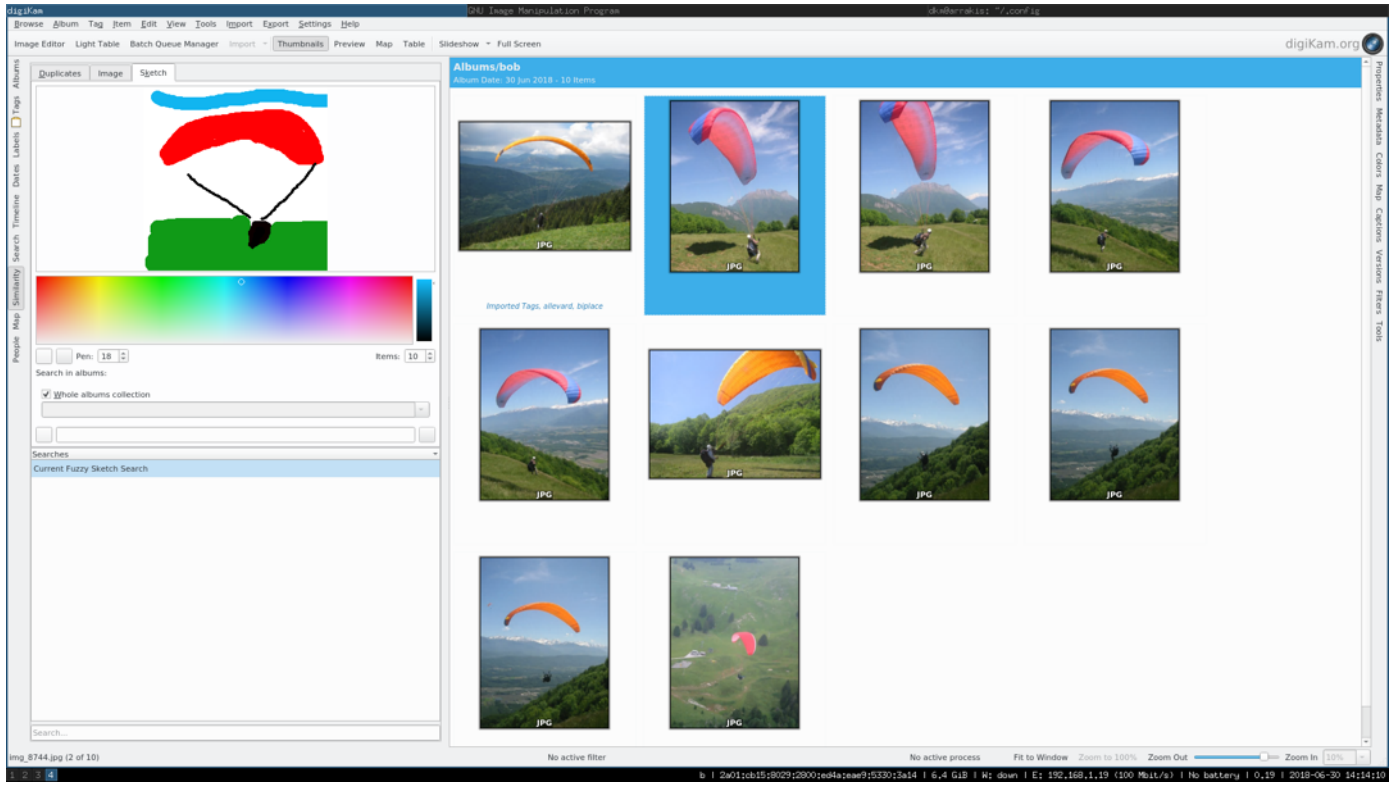


NAVER LABS Europe

Visual Search - Applications

Many applications

- Reverse Image search
  - Web search engine
  - Personal photo collection

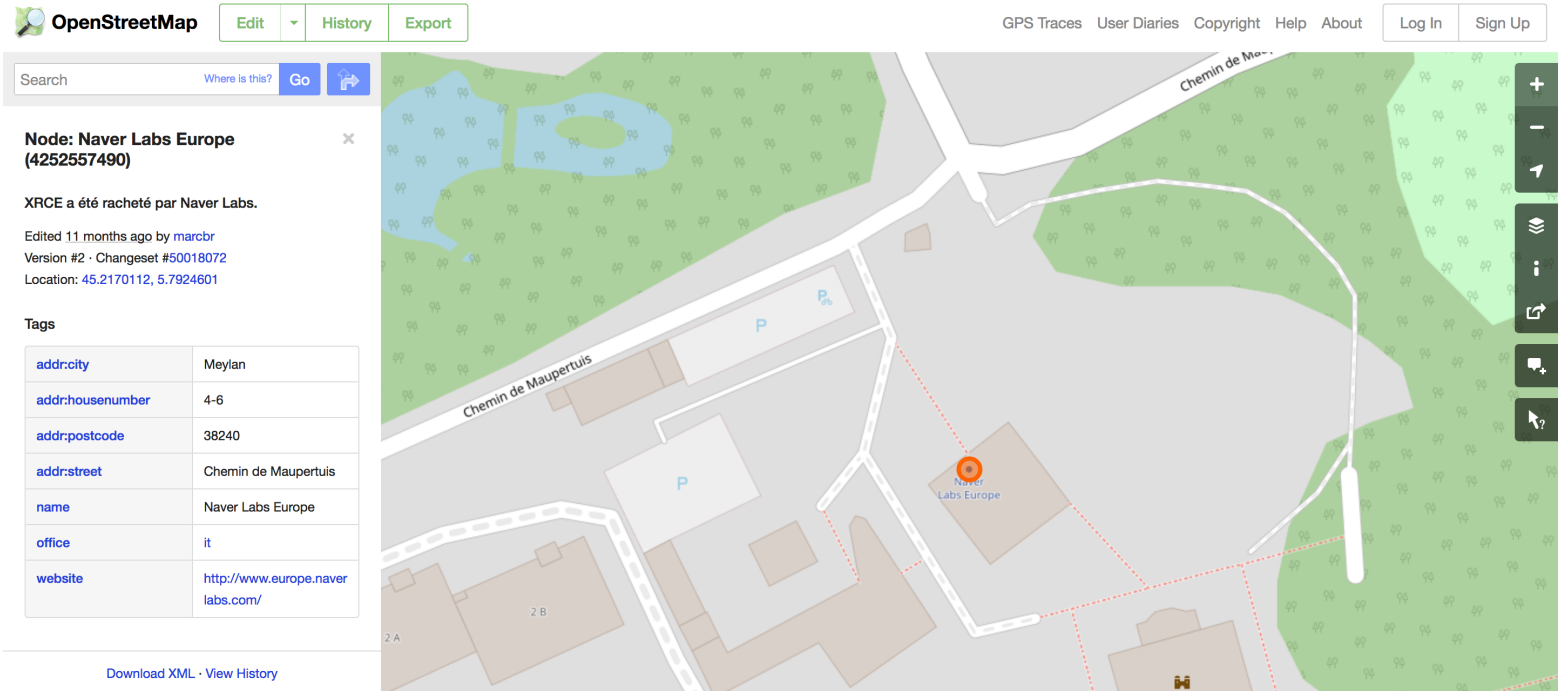


NAVER LABS Europe

Visual Search - Applications

Many applications

- Geolocalization

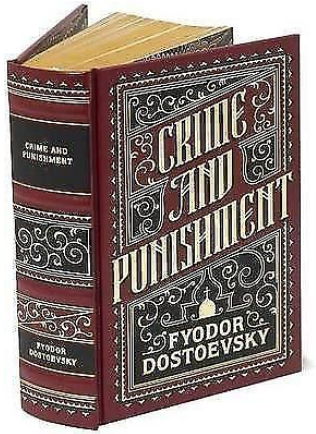


NAVER LABS Europe

Visual Search - Applications

Many applications

- Query for more information
  - Landmarks
  - Paintings
  - Movies
  - Book covers
  - Game covers
  - Packaged food



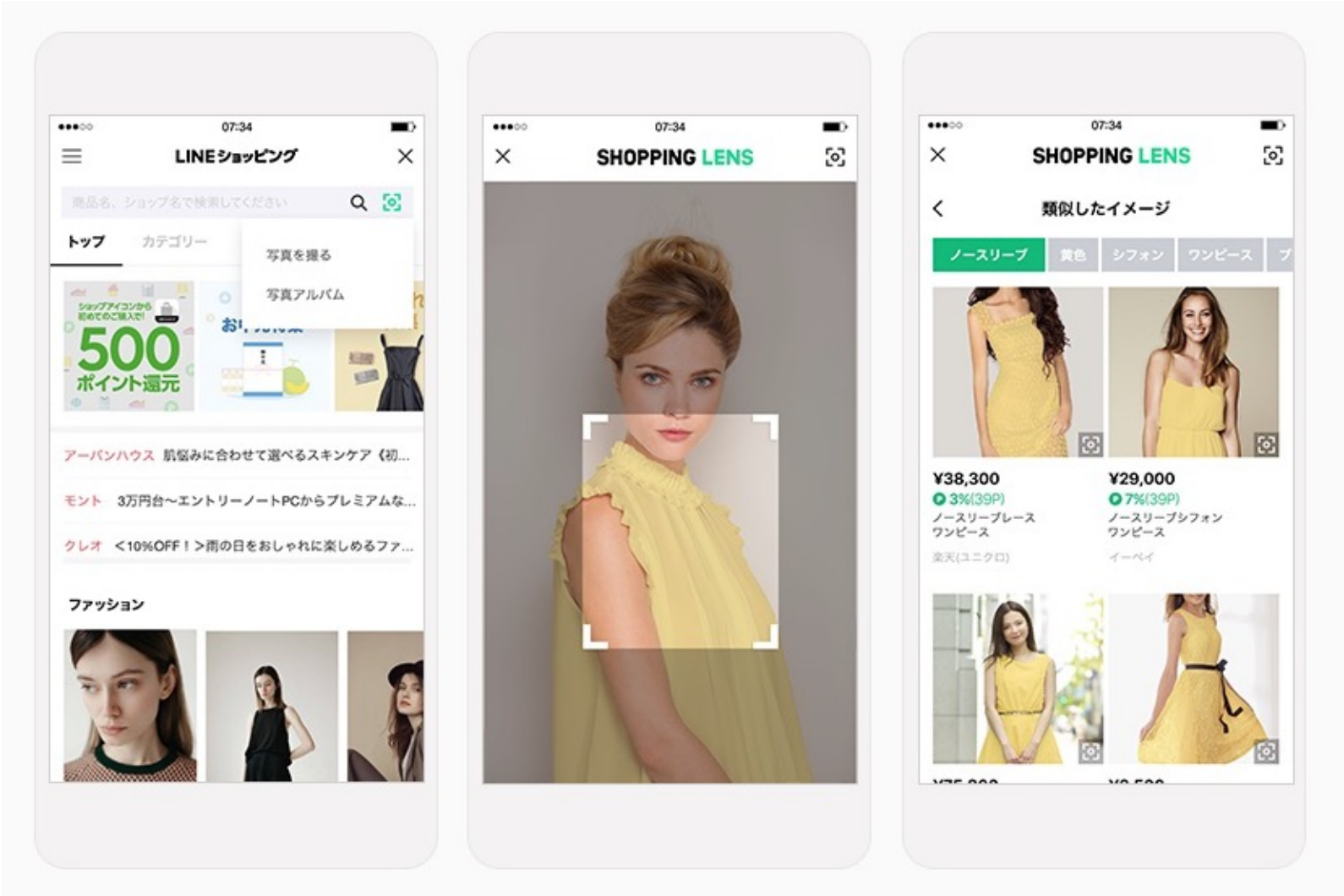
NAVER LABS Europe

Visual Search - Applications

NAVER

Many applications

- Shopping interfaces



NAVER LABS Europe

Visual Search - Applications

Many applications

- Ambient Intelligence

NAVER LABS Europe

Inherent ambiguity

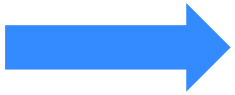
What can the user mean with such a single query?

NAVER LABS Europe

Inherent ambiguity

What can the user mean with such a single query?

Application dependent!



- Inject prior information
  - Hand-crafting detectors / descriptors to obtain some properties (repeatability, invariance, discrimination, compactness, etc.)
- Leverage training
  - Provided with training data, learn a descriptor appropriate for the task

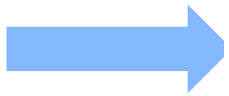
Outline

Object Search

Semantic Retrieval

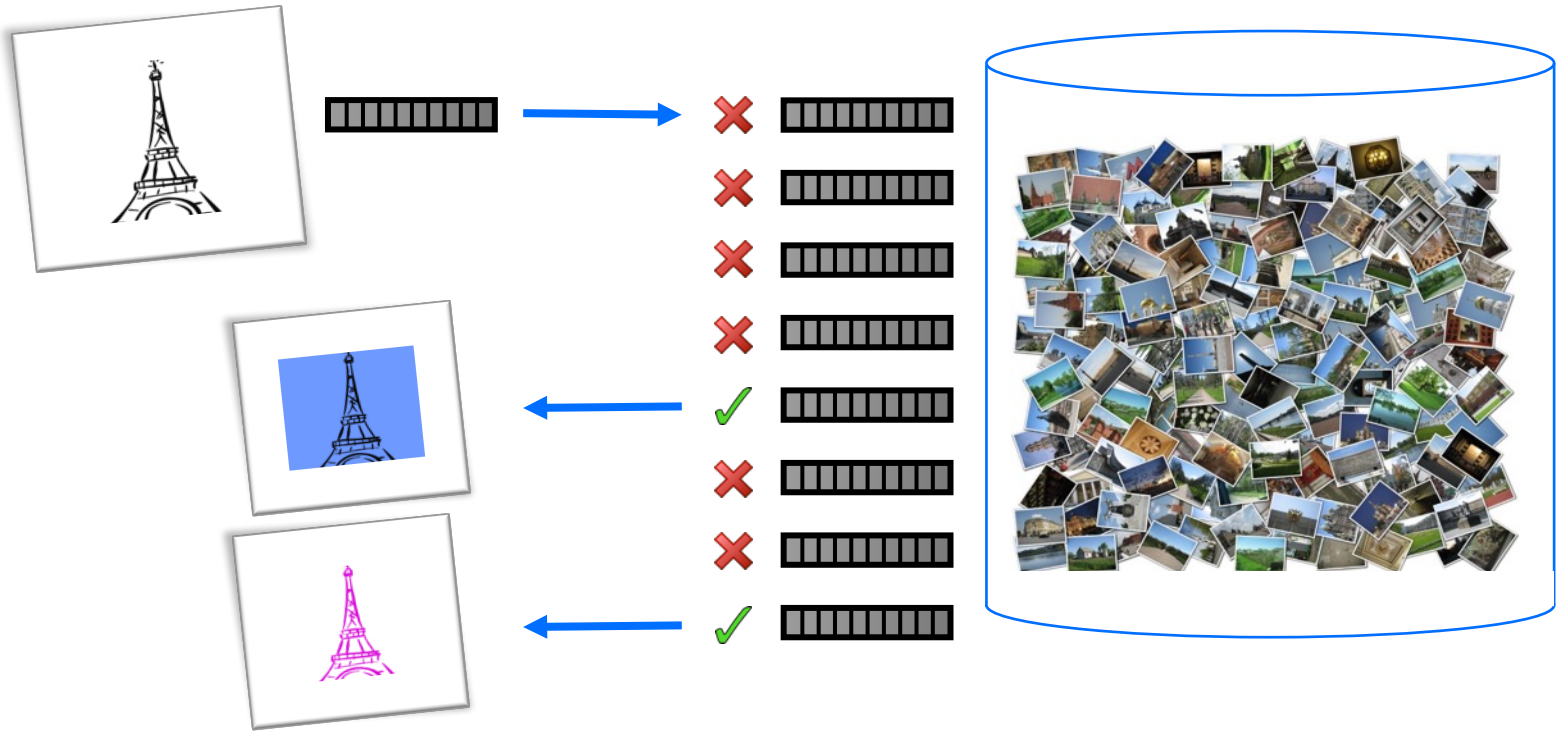
NAVER LABS
Europe

Object Search



NAVER LABS
Europe

Object Search *a.k.a.* ***Instance-Level Retrieval***



Families of representations

- **Early methods**
- Local representations
- Global representations
- Deep representations

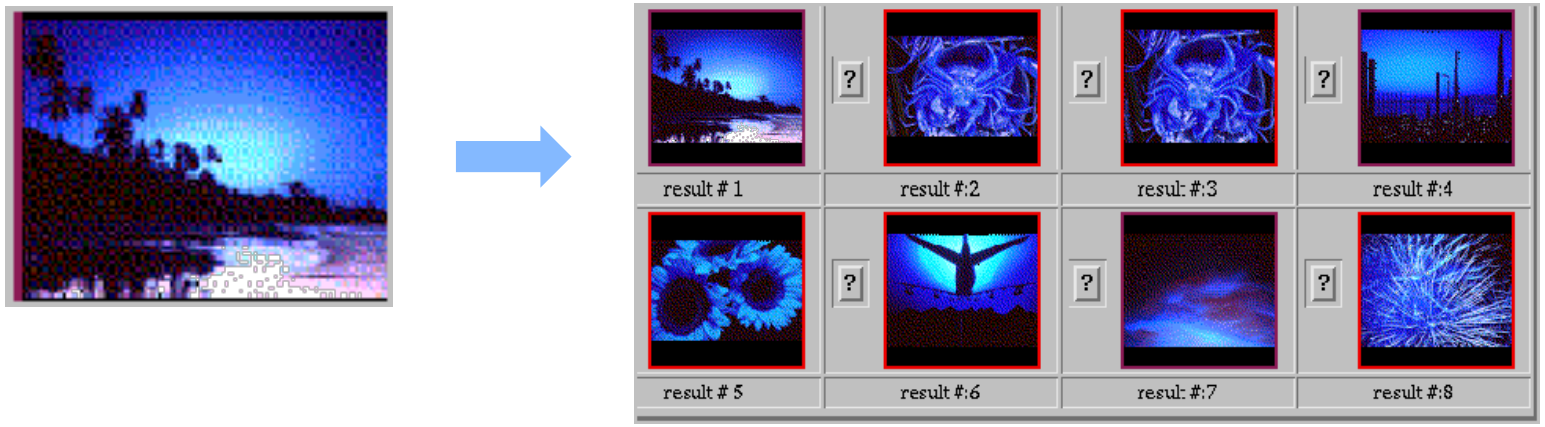
Corresponding similarity measures

NAVER LABS Europe

Early approaches

- Use low-level cues
  - Color histograms, texture and shape descriptors

[Swain & Ballard. IJCV91]



[Niblack et al.
The QBIC project. SPIE 93]

[Veltkamp et al. Content-Based Image Retrieval Systems: **a Survey**. 2000]

NAVER LABS Europe

Appearance variation

reference image

viewpoint

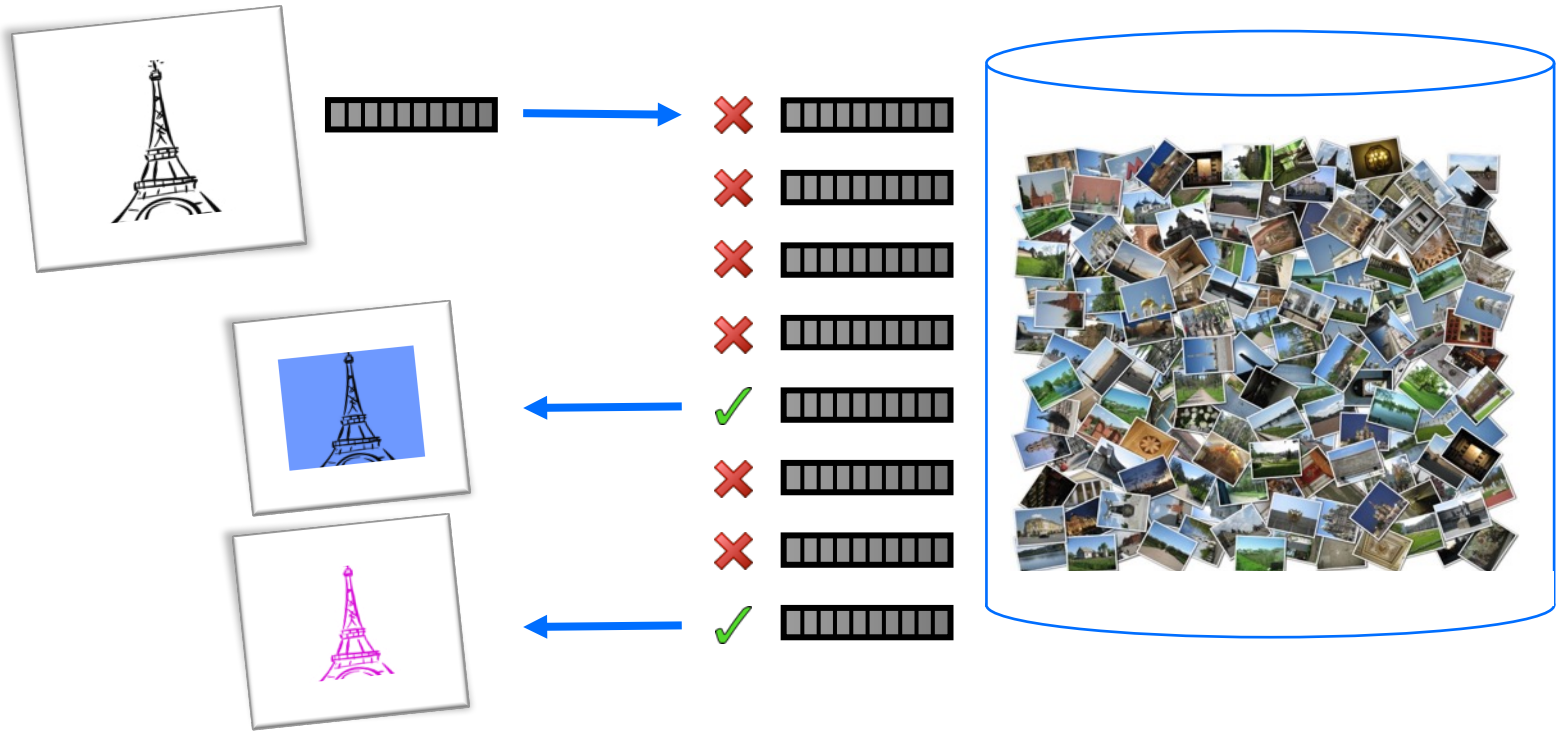
scale

occlusion

illumination

NAVER LABS Europe

Object Search *a.k.a.* ***Instance-Level Retrieval***



Families of representations

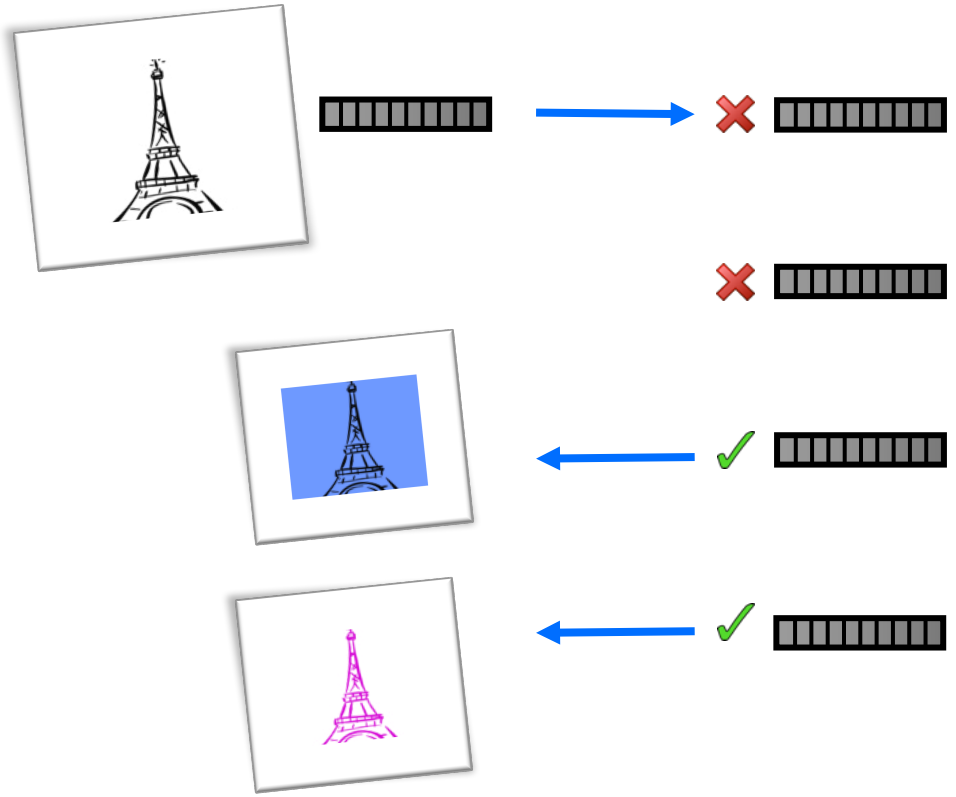
- Early methods
- **Local representations**
- Global representations
- Deep representations

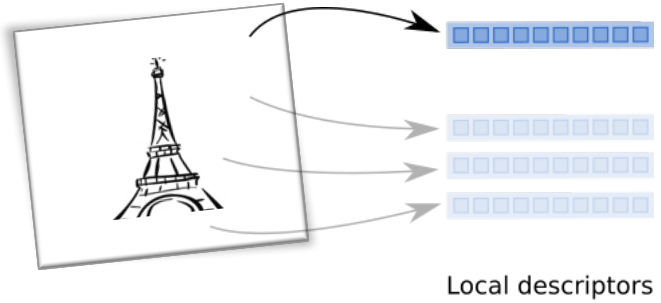
Corresponding similarity measures

NAVER LABS Europe

Local representations

It is challenging to capture all these invariances in a global representation

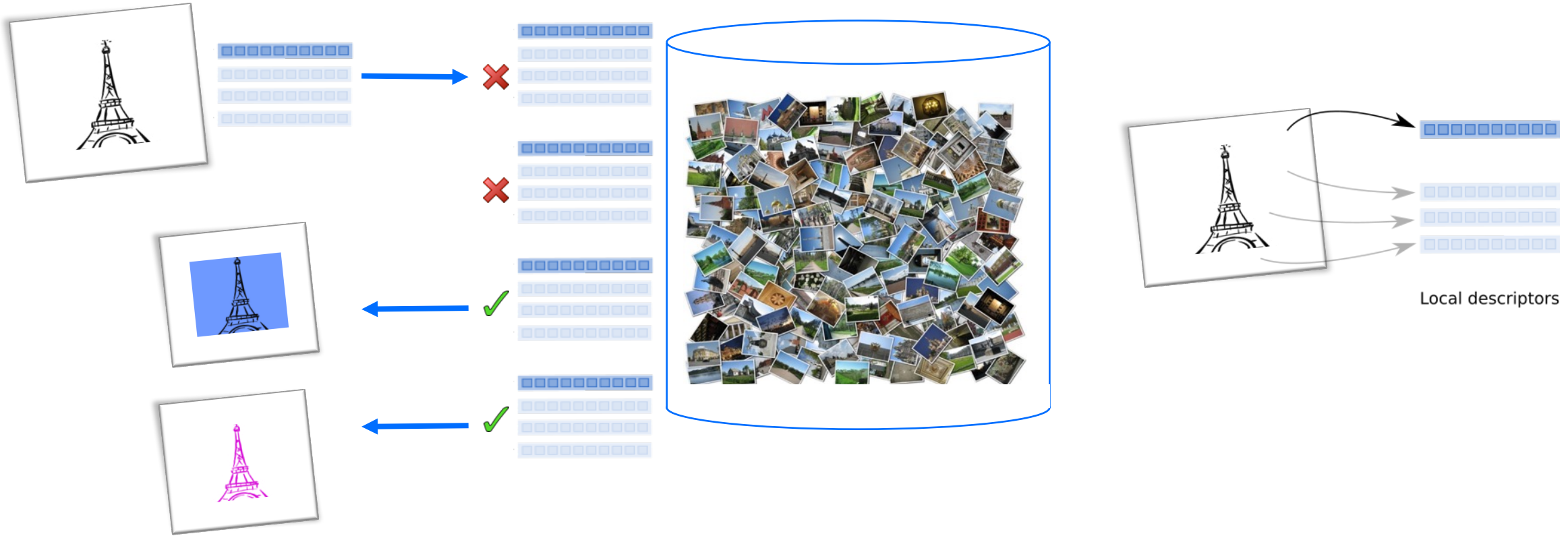




NAVER LABS Europe

Local representations

It is challenging to capture all these invariances in a global representation

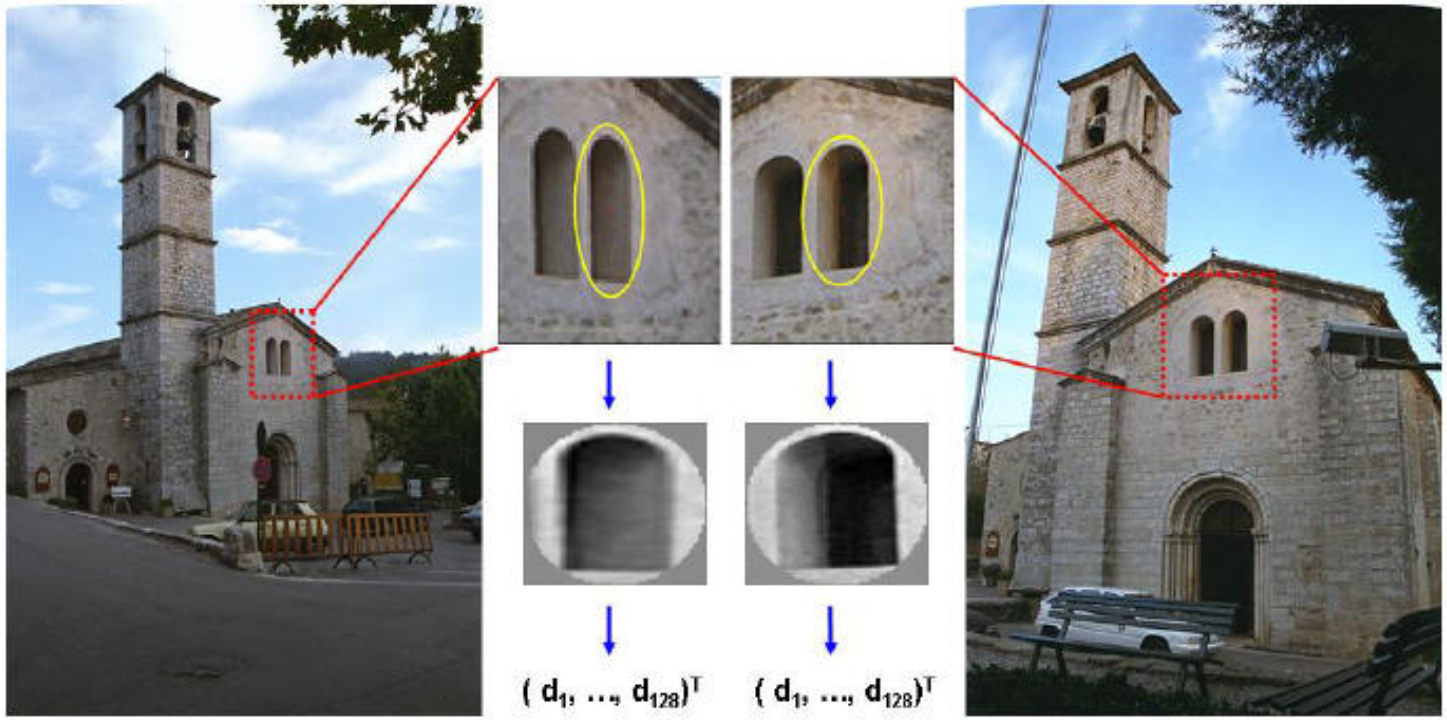


NAVER LABS Europe

Local representations

Local descriptors capture the local appearance:
invariant and discriminative

- Carefully chosen locations: repeatability



NAVER LABS Europe

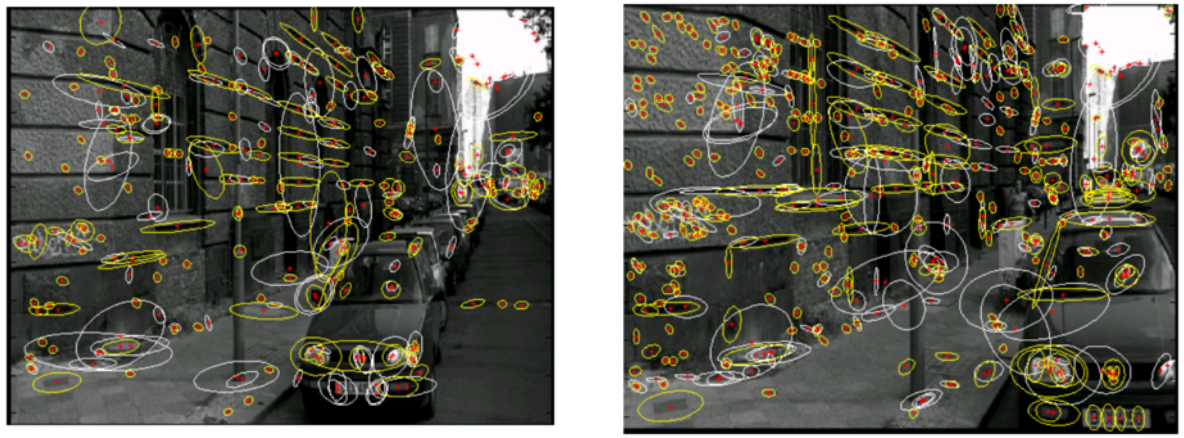
Local representations

Local descriptors capture the local appearance:
invariant and discriminative

- Carefully chosen locations: interest point detector

Interest point detectors

- Harris, Hessian, Hessian-Affine, MSER, etc.



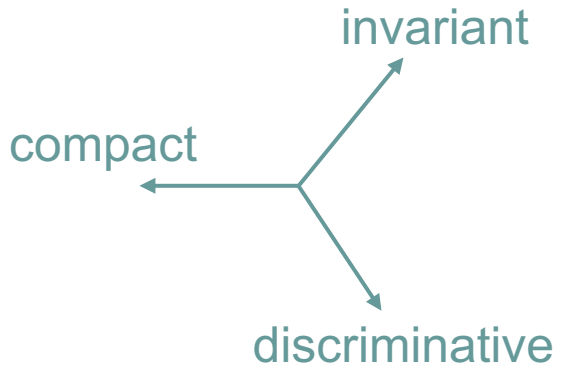
[Harris&Stephens1988, Mikolajczyk&Schmid2002, Matas et al, 2004]
Survey: [Mikolajczyk, IJCV 2005]

NAVER LABS Europe

Local representations

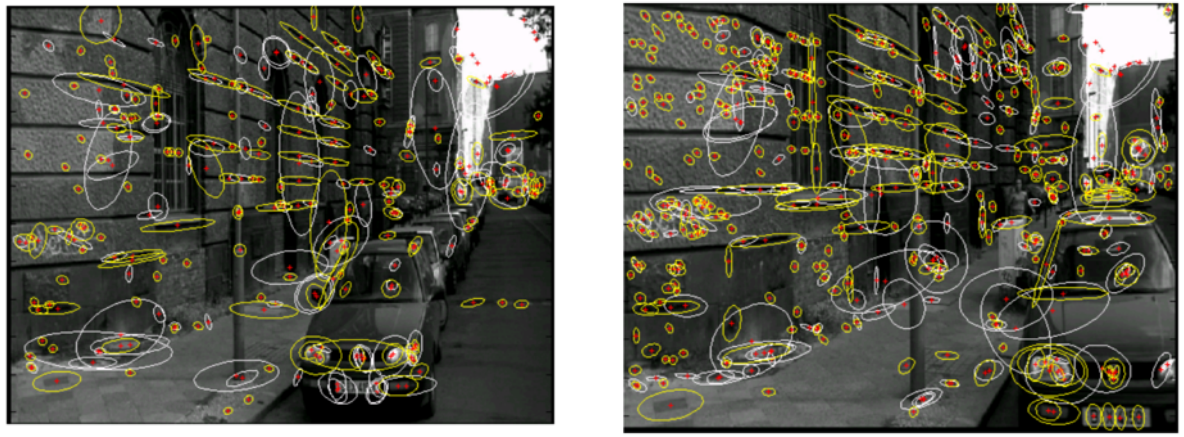
Local descriptors capture the local appearance: invariant and discriminative

- Carefully chosen locations: repeatability



Interest point detectors

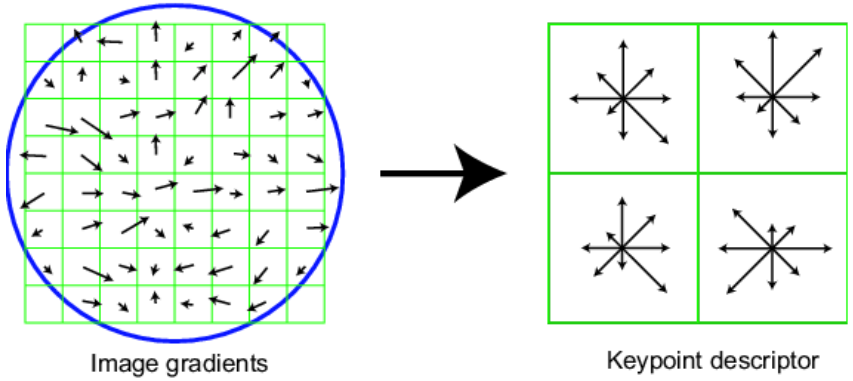
- Harris, Hessian, Hessian-Affine, MSER, etc.



[Harris&Stephens1988, Mikolajczyk&Schmid2002, Matas et al, 2004]
Survey: [Mikolajczyk, IJCV 2005]

Local descriptors

- SIFT, SURF, LBP, etc.

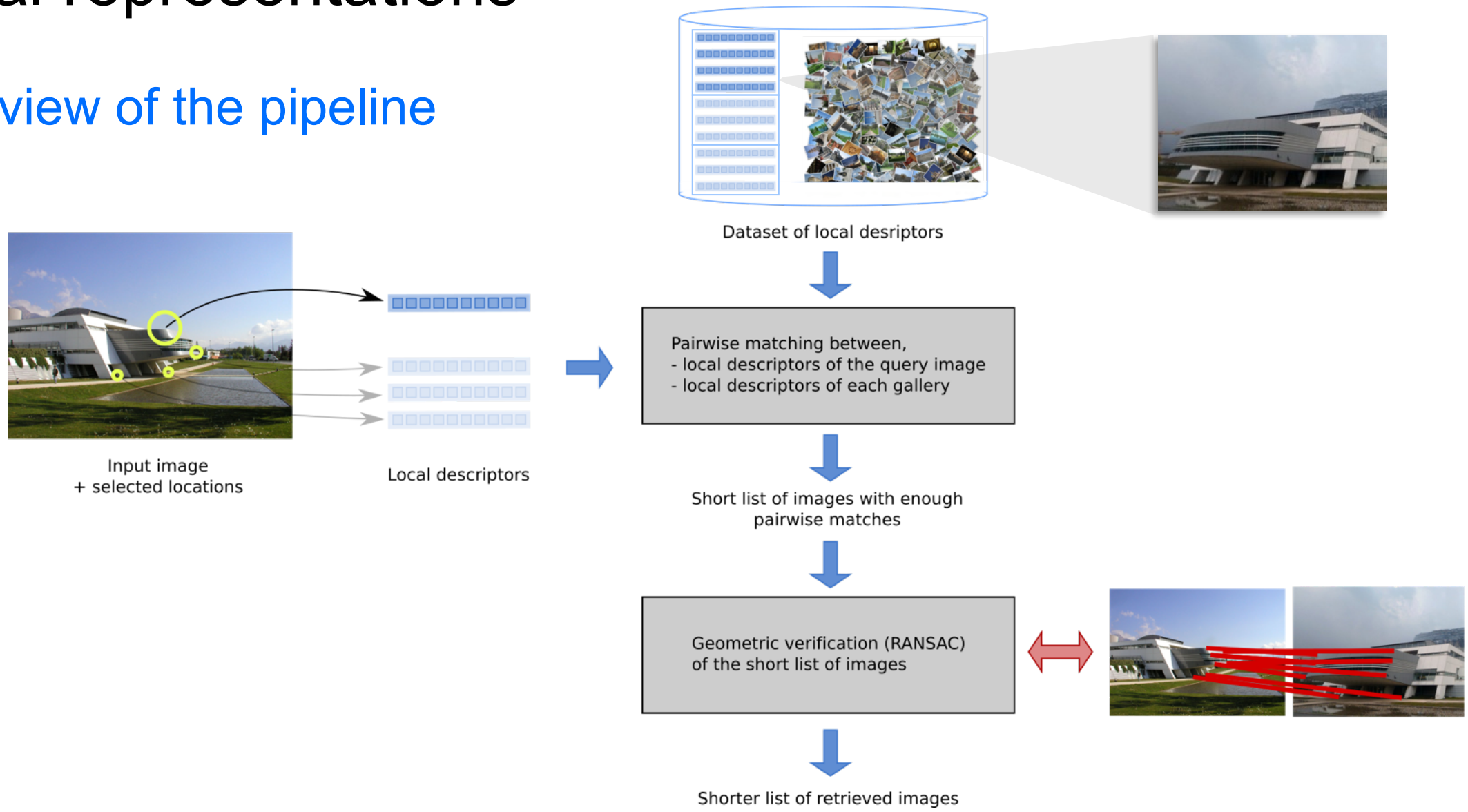


[Lowe IJCV04, Bay et al, 2008, Ojala et al, 2014]

NAVER LABS Europe

Local representations

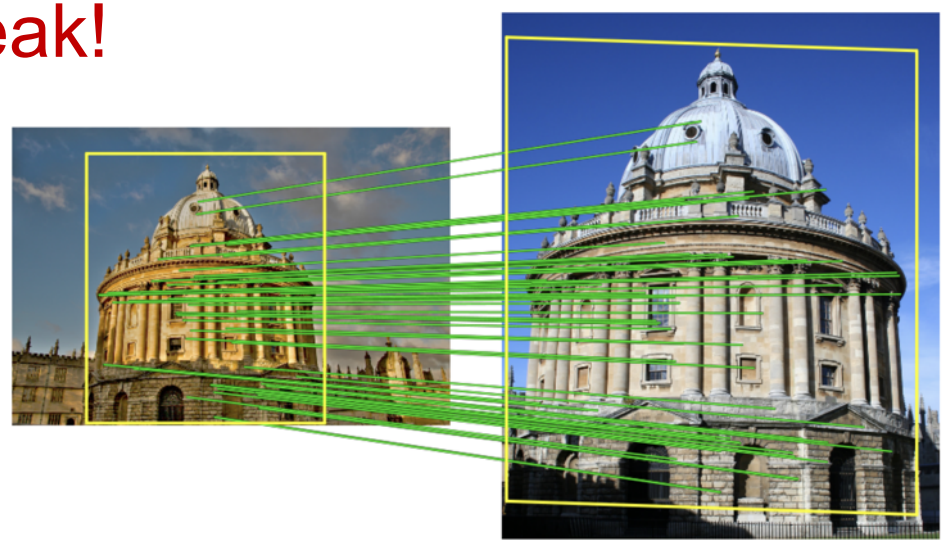
Overview of the pipeline



NAVER LABS Europe

Geometry verification

Pairwise local descriptor matching is too weak!



Geometrical verification

- **RANSAC** and improvements
  - from a set of local pairwise matches
  - find the geometric transformation that fits the highest number of matches
  - retain only images with enough matches for the estimated transformation
- Provides coarse object localization

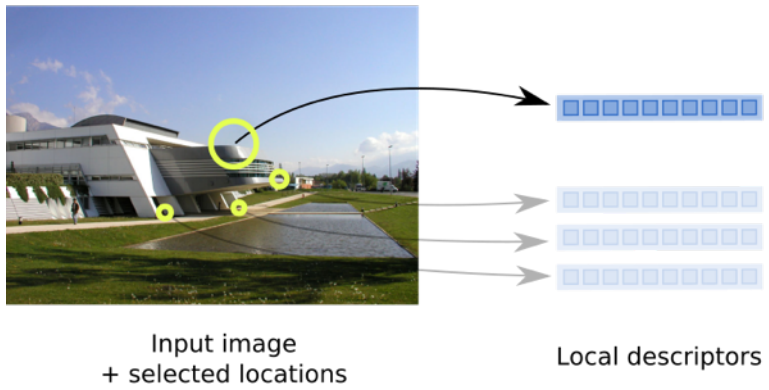
Costly process: verify only a short list of images

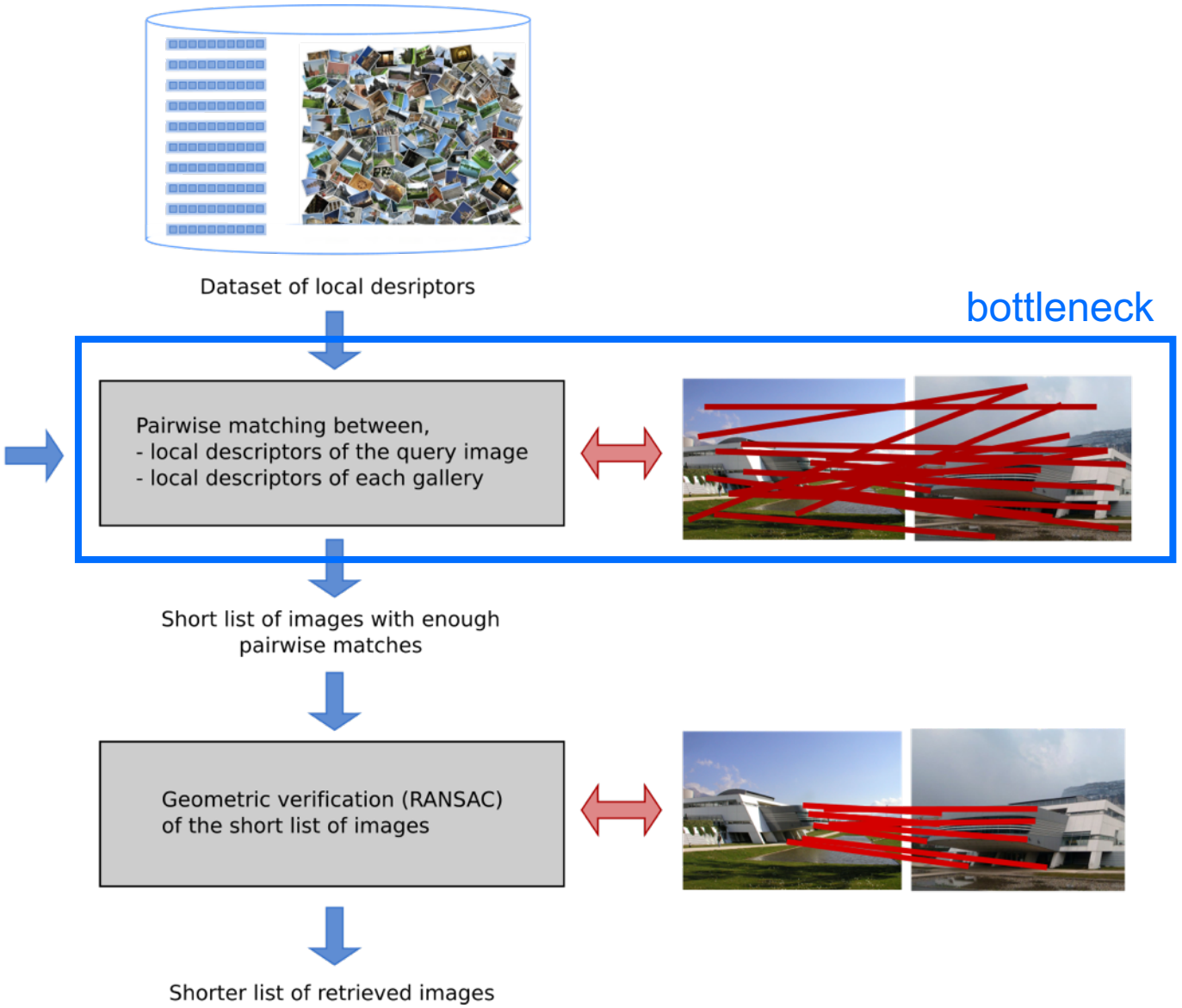
[Fischler & Bolles. 1981]
[Chum et al. 2007]
[Chum. PhD Thesis. 2005]

NAVER LABS Europe

Local representations

Overview of the pipeline





NAVER LABS Europe

Scaling the first selection process

1) Efficient approximate nearest neighbor search on local descriptors

- Randomized K-d trees and variants
- Locality-Sensitive Hashing (LSH)
  - Randomized hashing technique
  - Data-dependent variants: spectral hashing, semantic hashing

[Silpa-Anan & Hartley 2008,
Chum et al 2008, Muja & Lowe. 2009]

[Indik & Motwani, 1998, Weiss et al NIPS08,
Salakhutdinov & Hinton, SIGIR07]

Scaling the first selection process

1) Efficient approximate nearest neighbor search on local descriptors

- Randomized K-d trees and variants
- Locality-Sensitive Hashing (LSH)
  - Randomized hashing technique
  - Data-dependent variants: spectral hashing, semantic hashing

[Silpa-Anan & Hartley 2008, Chum et al 2008, Muja & Lowe. 2009]

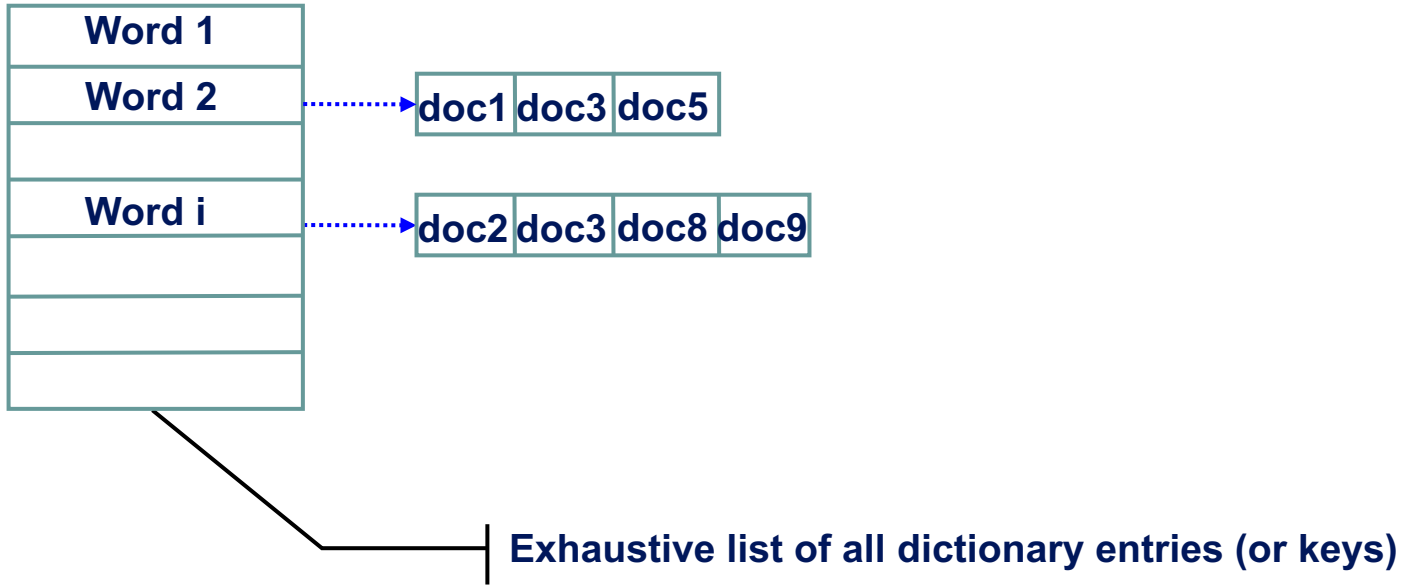
[Indik & Motwani, 1998, Weiss et al NIPS08, Salakhutdinov & Hinton, SIGIR07]

2) Quantize local descriptor space into a visual codebook

Let's take a moment to look at text retrieval

Inverted files for text retrieval

- This structure groups documents which have the same value for a given attribute
- It is typically used for text retrieval



- Querying for a word corresponds to querying for documents which contain that word
  - The cost is proportional to the number of documents to retrieve

Textual document retrieval

text retrieval

- Example
  - Vocabulary = {cat, dog, sheep, fish}
  - Representation space: $\mathbb{R}^4$
  - Sentence: "It is raining cats and dogs"

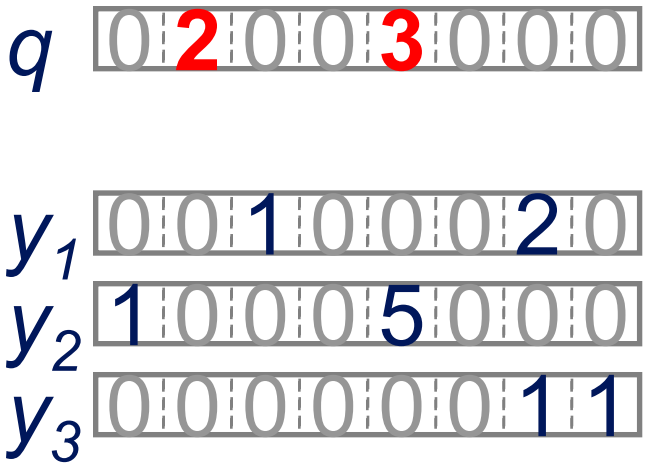
Sentence representation: $[1/4,1/4,0,0]^t$

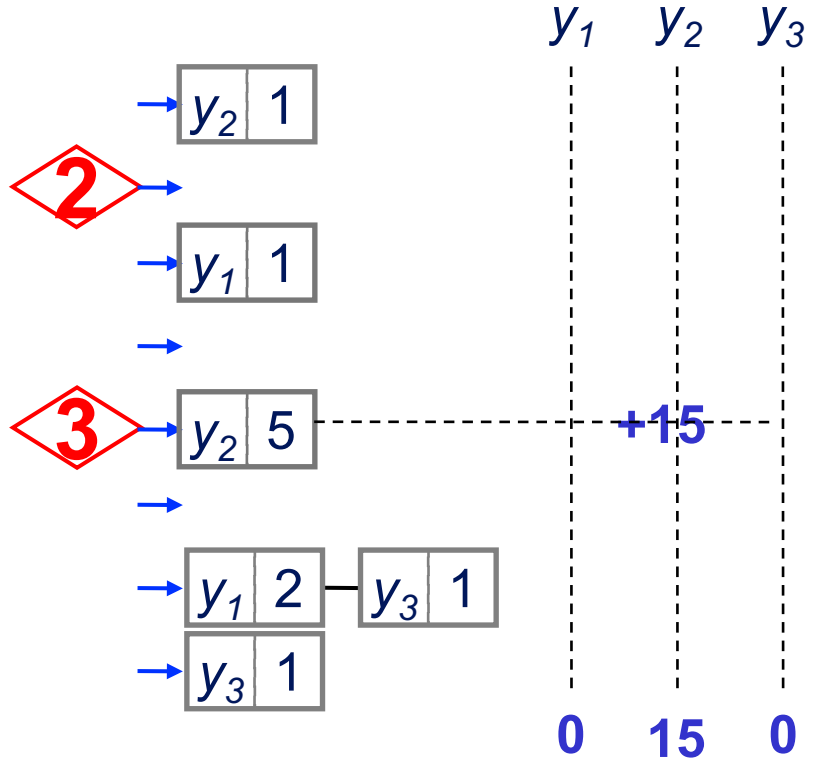
- Task: document retrieval

  = find the most similar similar document representation vectors to the vector q which represents the query document
  - Most similar requires a similarity / dissimilarity measure
  - Typically: similarity is computed as a dot-product

Inverted files: distance between two sparse vectors

- We assume a vocabulary of 9 words
- We pre-compute the descriptors for the 3 elements of the database Y
- We compute the descriptor of the query q on the fly
- Query q & elements from the database Y are sparse vectors
- Inverted file: efficiently compute the scalar





NAVER LABS Europe

Visual search = Documents are images

Back to image retrieval

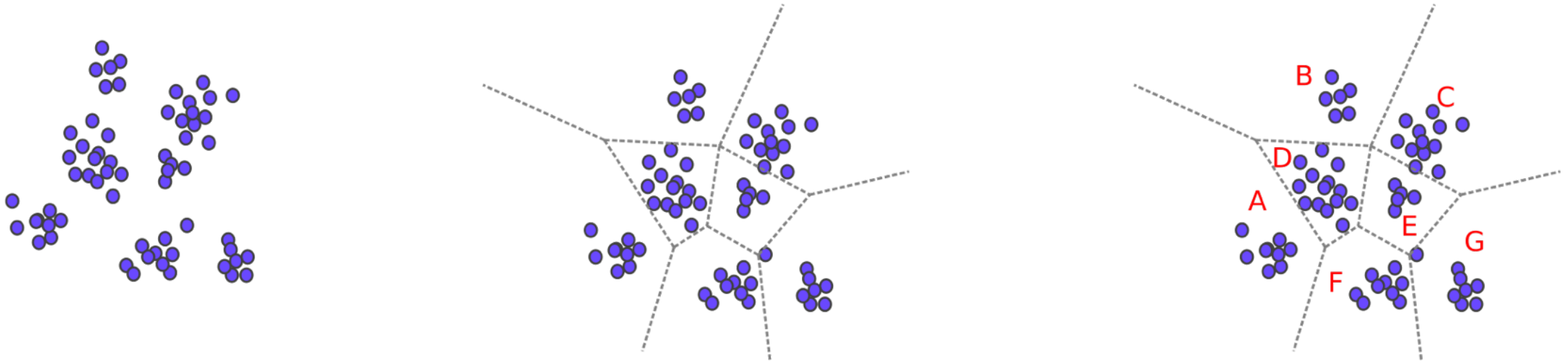
- Problem: the feature space of local descriptor is continuous (e.g. the 128-dimensional SIFT representation)
  - There potentially exists an infinity of visual features
  - There is no such concept as a vocabulary as a uniquely defined set of ”words” for image, like it is the case for text

- Solution: create visual vocabulary!!
  - Also called *visual codebook*
  - Create it specifically for a given domain / dataset
  - We construct it by quantifying the space of local descriptors

Quantization of visual descriptors: principle

Principle:

- Discretize the local descriptor space
- 2 descriptors match i.i.f. they fall in the same bin



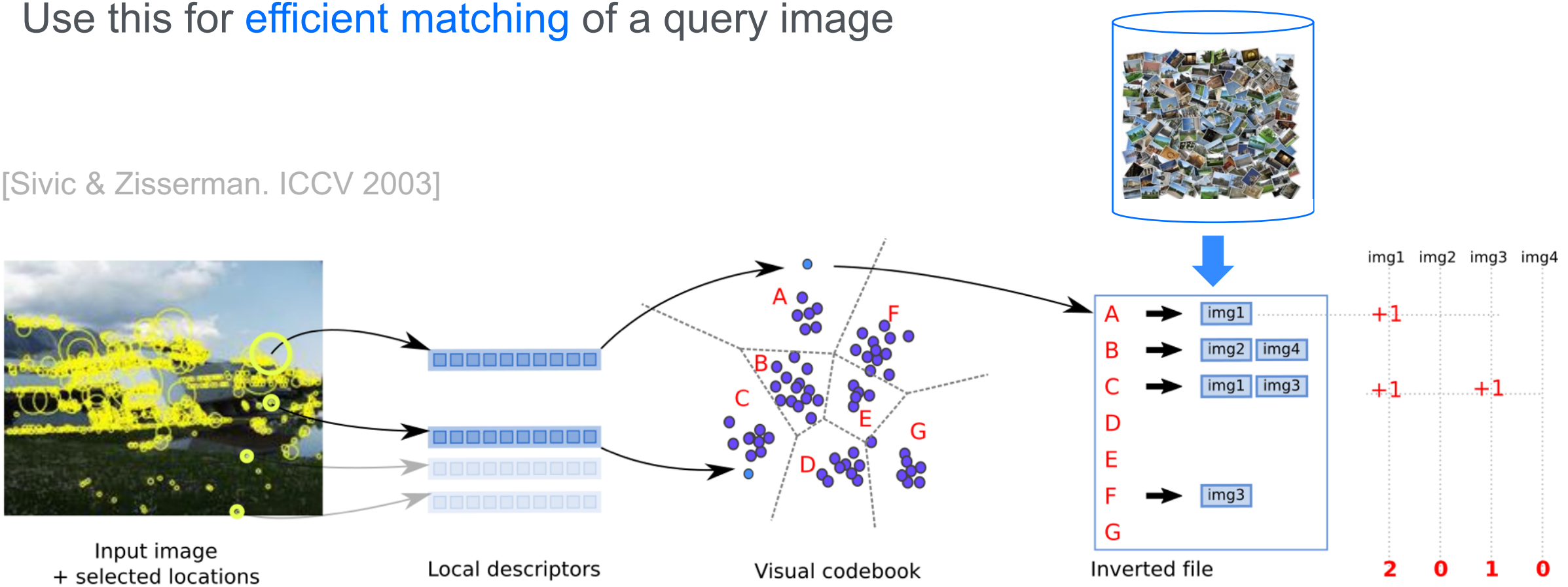
NAVER LABS Europe

Leverage quantization for efficient retrieval

Create an inverted file with all (quantized) descriptors of the gallery set

Use this for efficient matching of a query image

[Sivic & Zisserman. ICCV 2003]



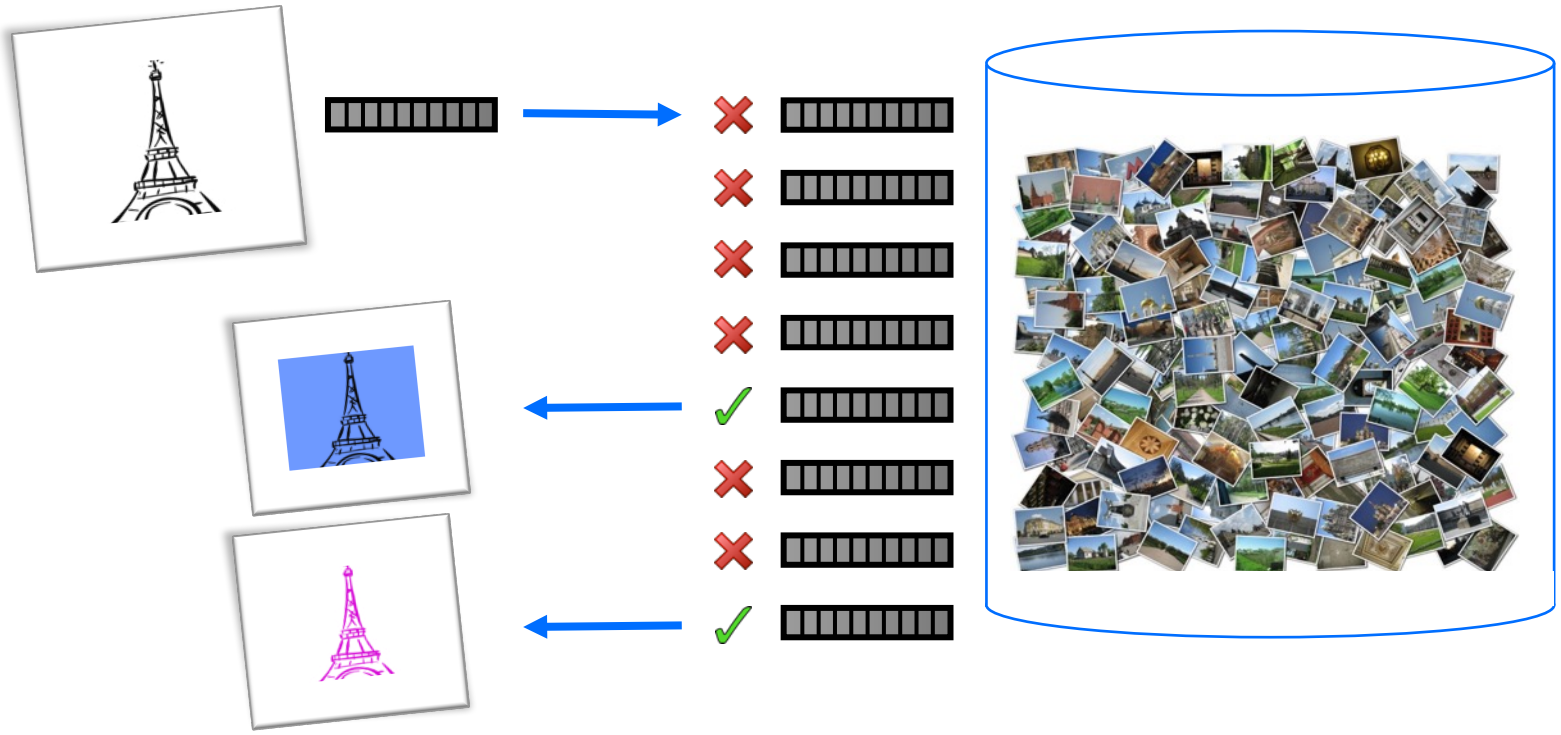
NAVER LABS Europe

Local description of images: summary and limitations

- **Local representations**
  - **In practice:** this approach gives very good results
  - Main limitations:
    - It requires to store all these local descriptors
      → it is very memory expensive!
    - It requires to match all the descriptors of the query image vs all descriptors of all images of the potentially very large dataset
      → it is very computationally expensive, hence very slow!
    - Because the geometrical verification is very slow
      → we can only use it on a small-subset of images selected a priori as more relevant

- **Coming next: alternative approach**
  - Using global descriptors which aggregate individual local descriptors in a single global descriptor per image
    - Local descriptors are not stored anymore, only an aggregated descriptor is
  - Computing the similarity between image pairs is obtained directly by computing global descriptors

Object Search *a.k.a.* ***Instance-Level Retrieval***



Families of representations

- Early methods
- Local representations
- **Global representations**
- Deep representations

Corresponding similarity measures

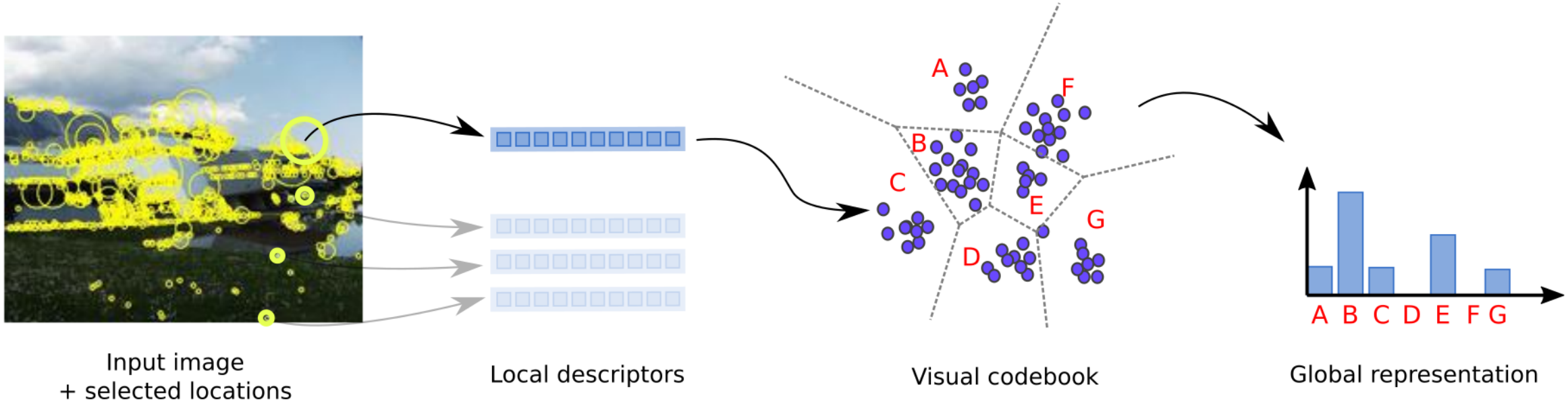
NAVER LABS Europe

From quantization to bag-of-visual-features

Principle

[Sivic & Zisserman. ICCV 2003]
[Csurka et al. ECCV SLCV 2004]

- Extract local descriptors
- Convert local descriptors into visual words, using a visual codebook
- Represent images as a histogram of occurrences



NAVER LABS Europe

How can we refine this description?

Relatively coarse representation

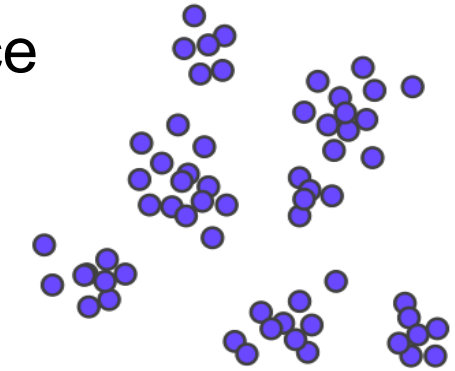
Solution 1: more entry in the codebook

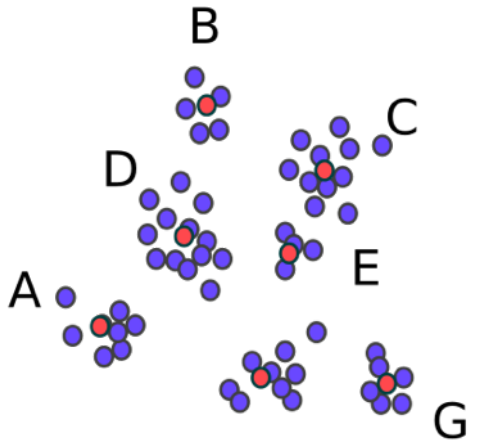
- drawback: significant computational cost

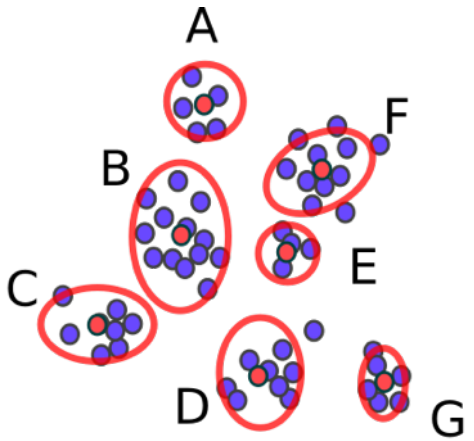
[Li et al. ICCV 2009]
[Yang et al. ECCV 2010]

Solution 2: beyond counting, adding higher order statistics

- Mean
- Variance





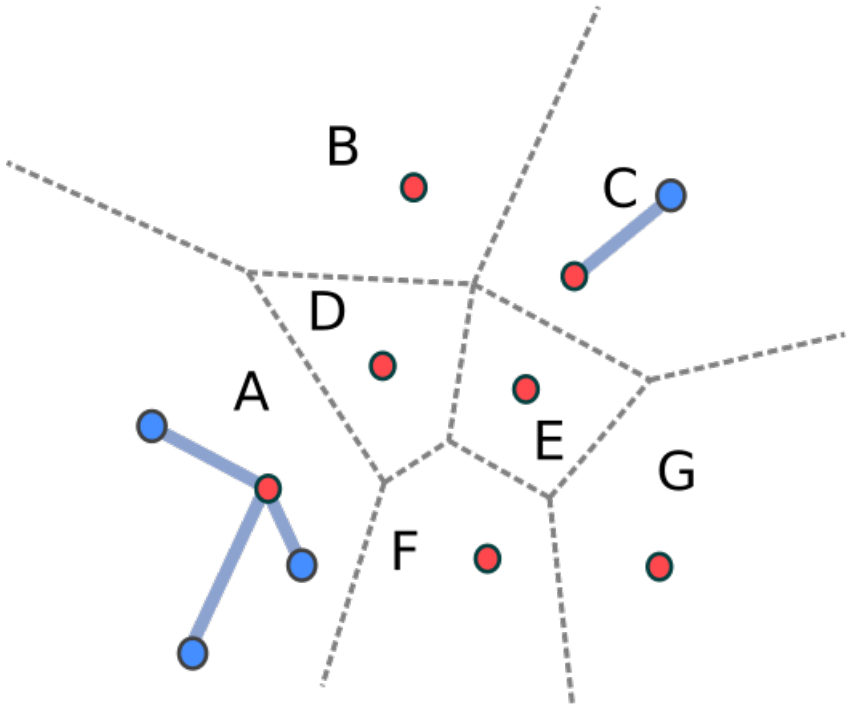


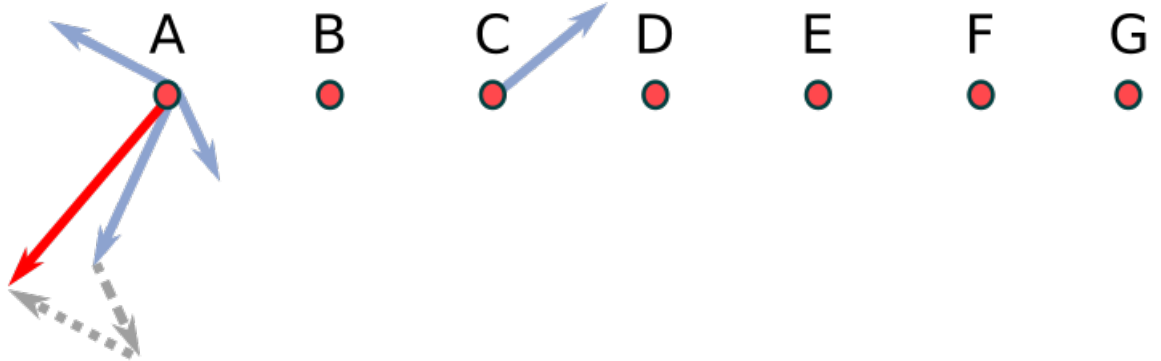
NAVER LABS Europe

Extension to higher order statistics

Mean: VLAD (Vector of Locally Aggregated Descriptors)

- Aggregate all descriptors assigned to the same visual word





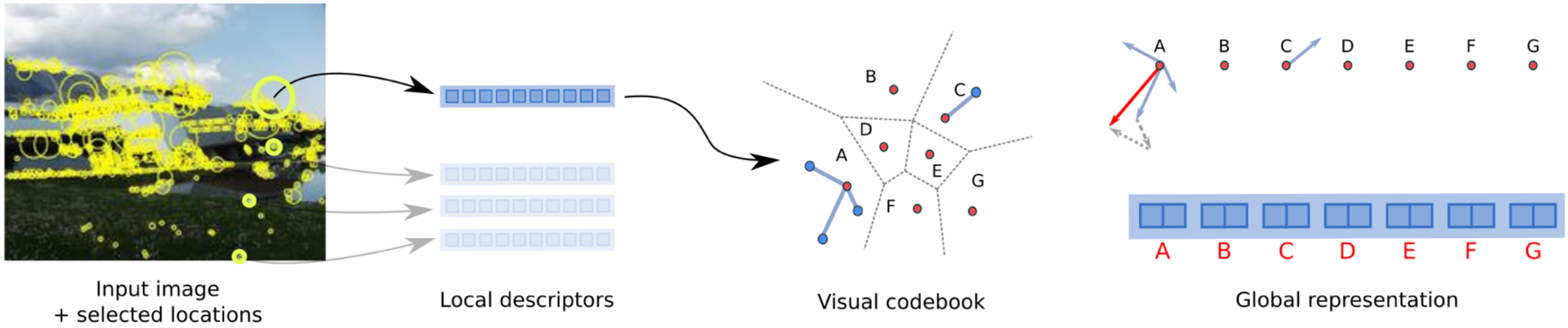
[Jégou et al. CVPR 2010]

NAVER LABS Europe

Extension to higher order statistics

Mean: VLAD (Vector of Locally Aggregated Descriptors)

- Aggregate all descriptors assigned to the same visual word
- Concatenate vectors for individual words



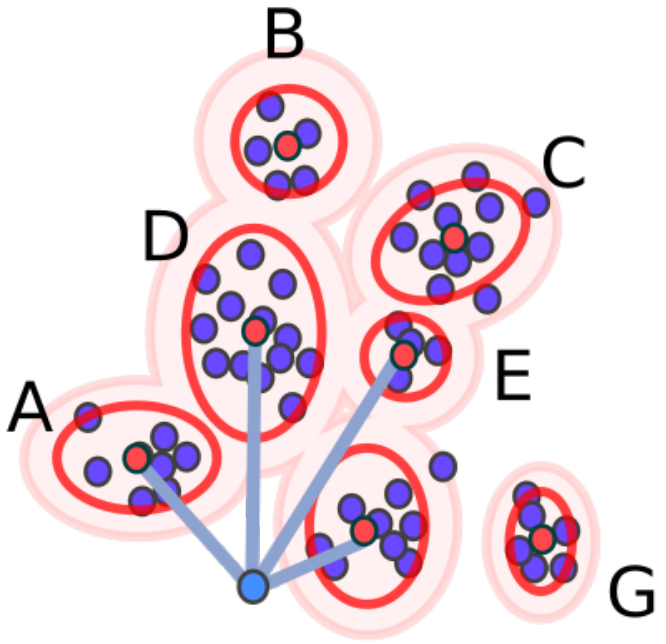
[Jégou et al. CVPR 2010]

NAVER LABS Europe

Extension to higher order statistics

Mean + Variance: Fisher-Vector

- Probabilistic codebook as a mixture of Gaussians
- Descriptors soft-assigned to words
- Compute the gradient of the log-likelihood w.r.t. the parameters of the model



[Perronnin and Dance. CVPR07]
[Perronnin et al. CVPR10]

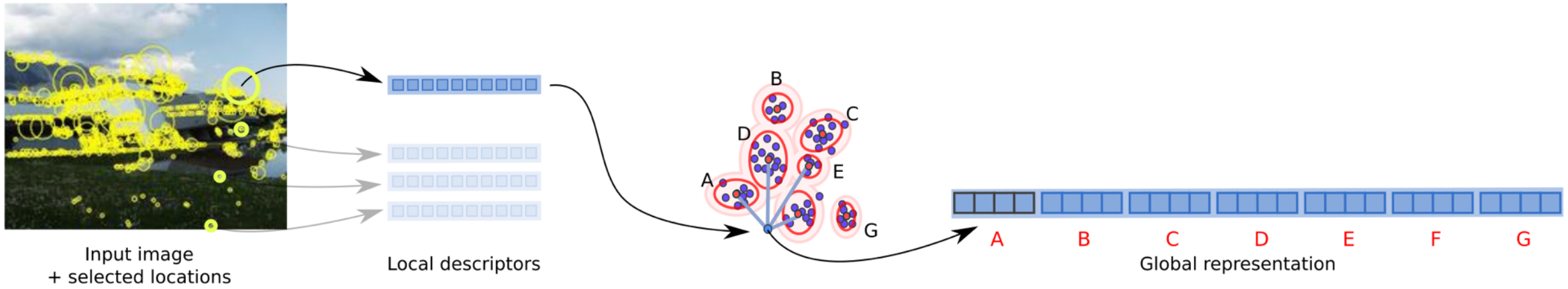
NAVER LABS Europe

Extension to higher order statistics

Mean + Variance: Fisher-Vector

- Aggregate the contribution of all local descriptors
- Concatenate statistics for all the visual words

[Perronnin and Dance. CVPR07]
[Perronnin et al. CVPR10]



NAVER LABS Europe

Evaluation on standard benchmarks

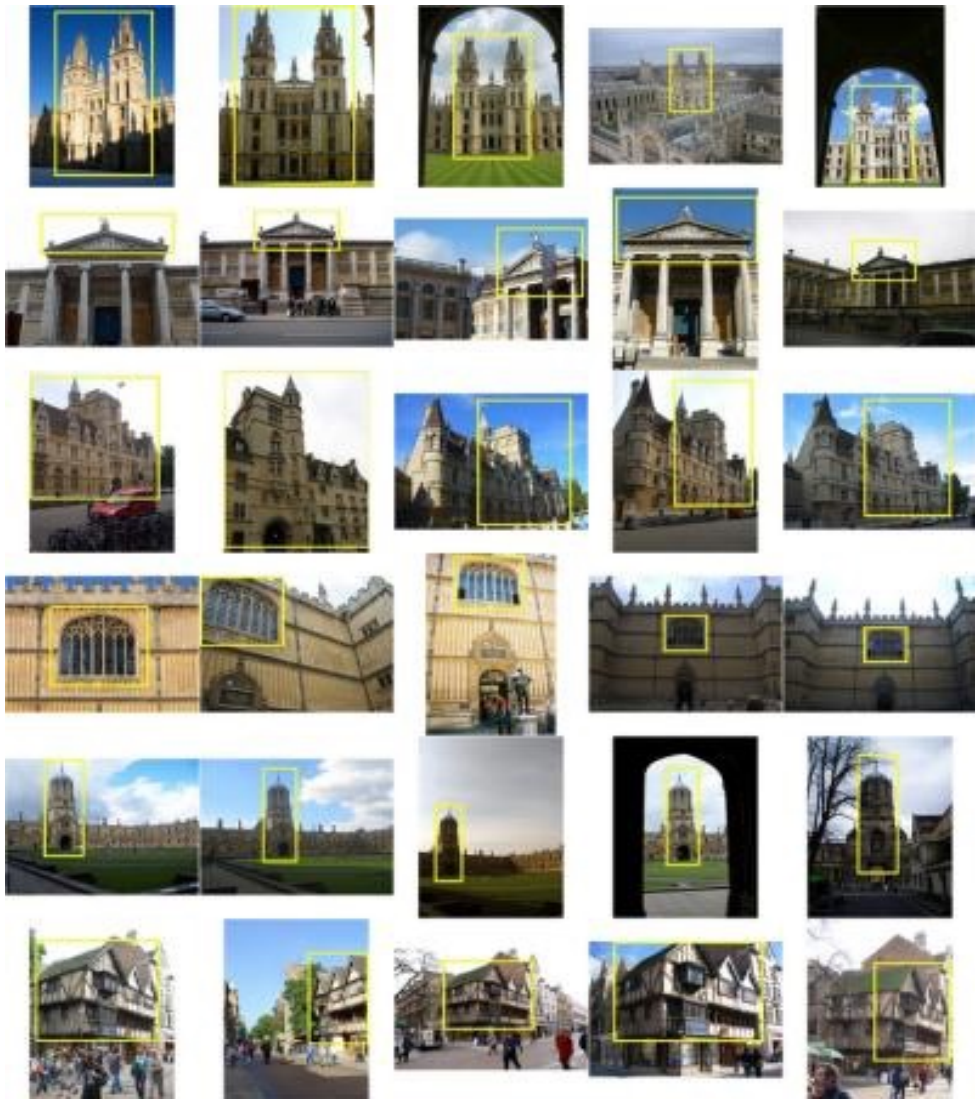
Oxford dataset

- 5,000 images
- 55 queries
- 11 landmarks

Evaluation

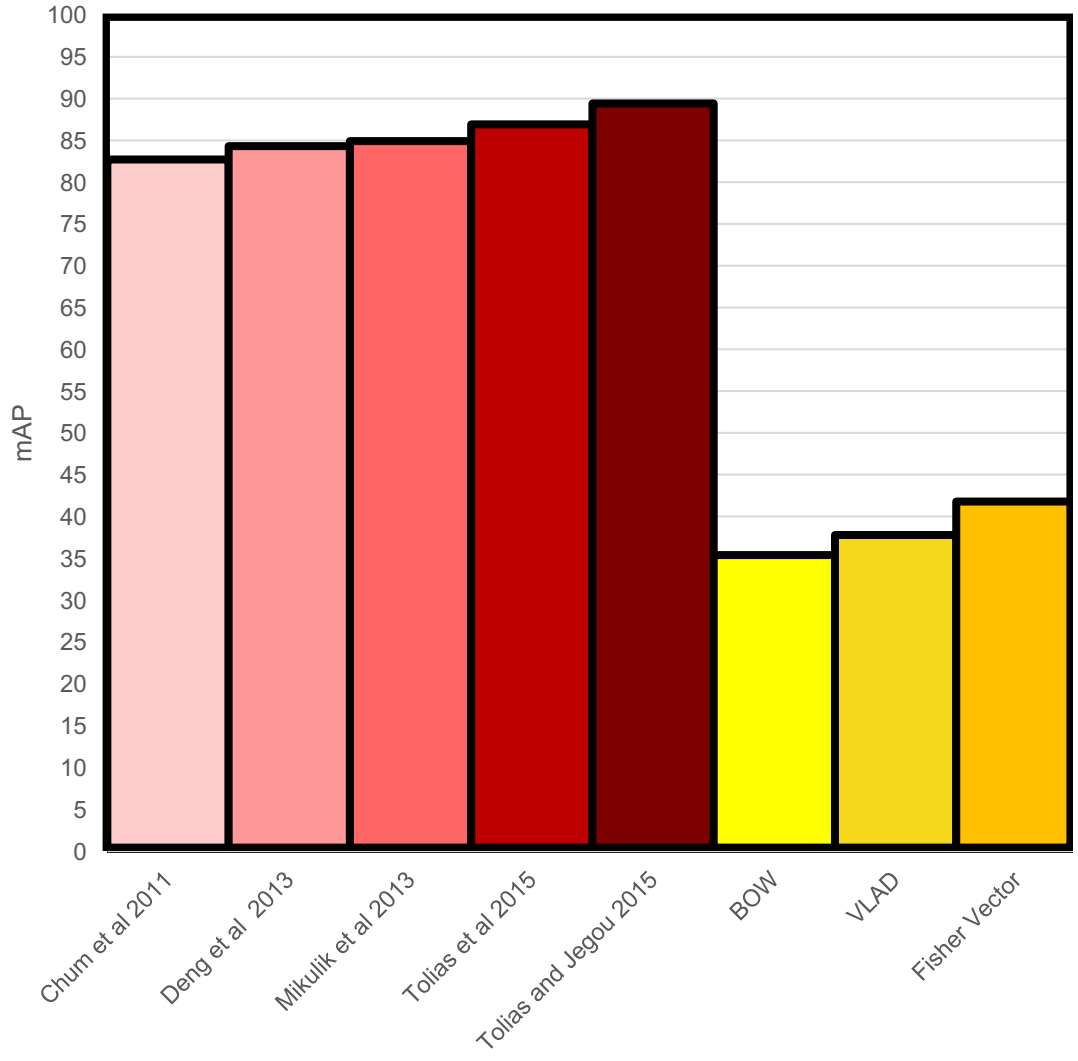
- mean Average Precision (mAP)

[Philbin et al. CVPR07]



NAVER LABS Europe

Experiments



Oxford 5K

Local representations

Global representations

NAVER LABS Europe

Standard approaches - Summary

Matching-based methods – local representations

- highest accuracy
- but high cost of matching and geometry verification

[Philbin et al, 2007, Chum et al 2007, Chum et al 2011, Jegou et al, 2008, Chum et al, 2009, Deng et al, Mikulik et al 2013, Tolias et al 2015, Tolias and Jegou 2015,]

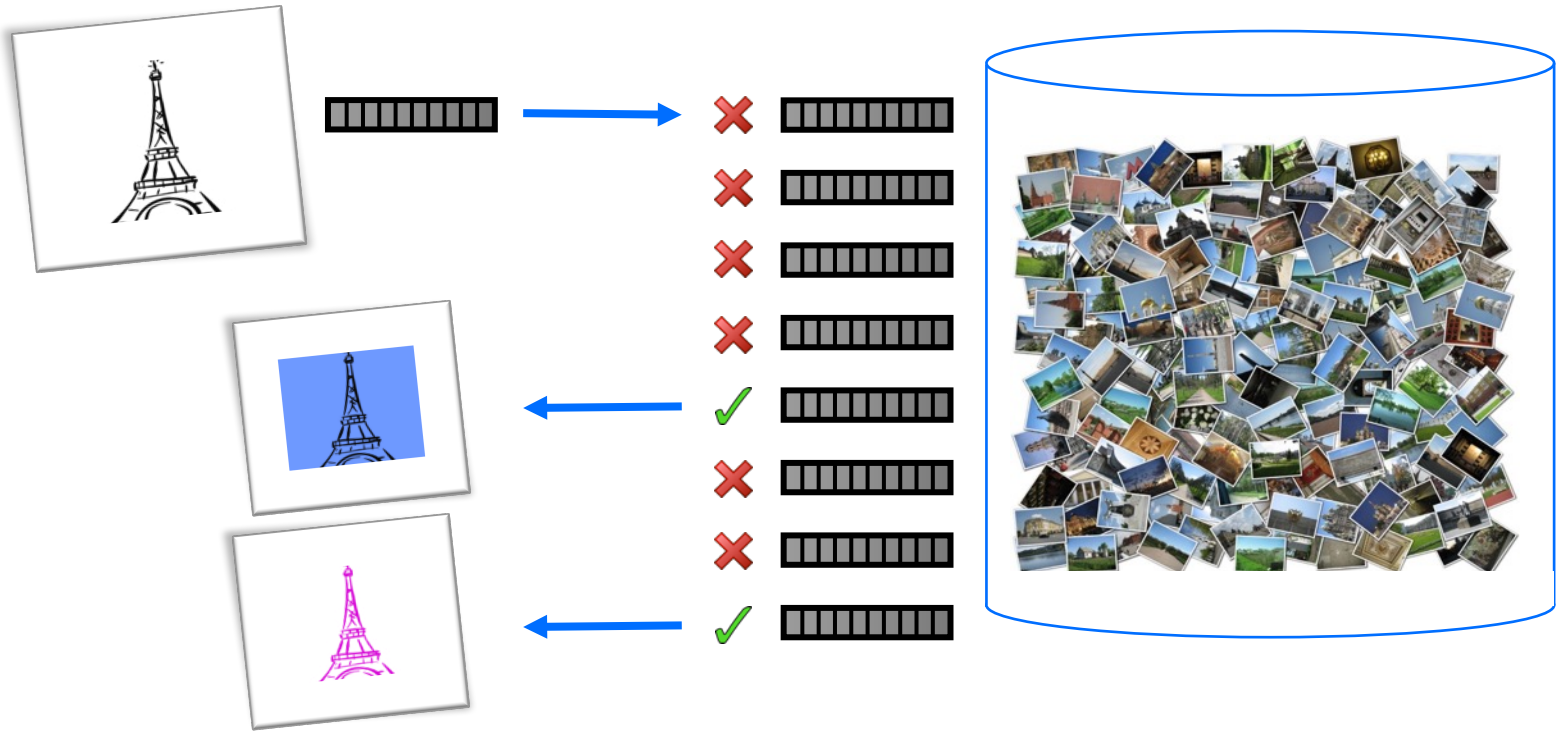
Aggregation methods – global representations

- faster and more efficient
- but lower accuracy than matching ones

[Jegou et al. CVPR 2010, Perronnin et al. CVPR 2010]

NAVER LABS Europe

Object Search *a.k.a.* ***Instance-Level Retrieval***



Families of representations

- Early methods
- Local representations
- Global representations
- **Deep representations**

Corresponding similarity measures

NAVER LABS Europe

First deep learning approaches

Principe

- Train a neural network (e.g. a CNN) on a classification task
  - For instance for the classification of 1000 classes of ImageNet
- Use this neural network as a “black box” to extract image representations
  - For instance the output of a layer (typically the one before the last one)
- Optionally, normalize representations
- Compare them with scalar products

Motivation

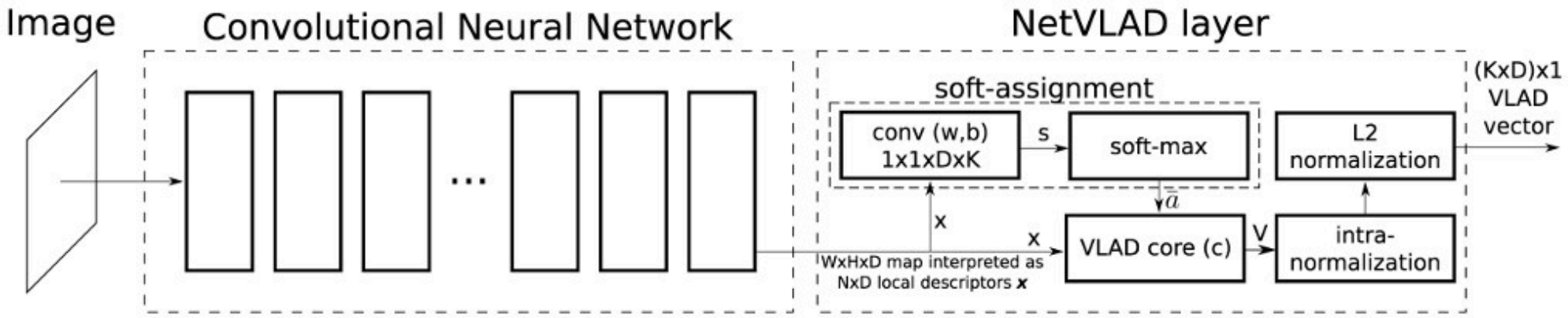
- A good classification model should be able to capture semantic information and should allow to represent images for retrieval task as well

Combining CNN and VLAD: **NetVLAD**

Principle:

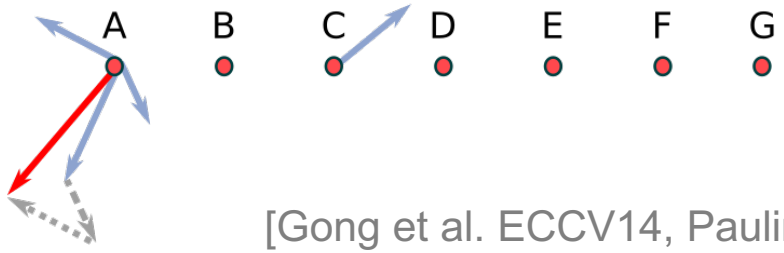
- VLAD block at the end of CNN

[Arandjelovic et al. CVPR16]



Input image

CNN



[Gong et al. ECCV14, Paulin et al ICCV15]

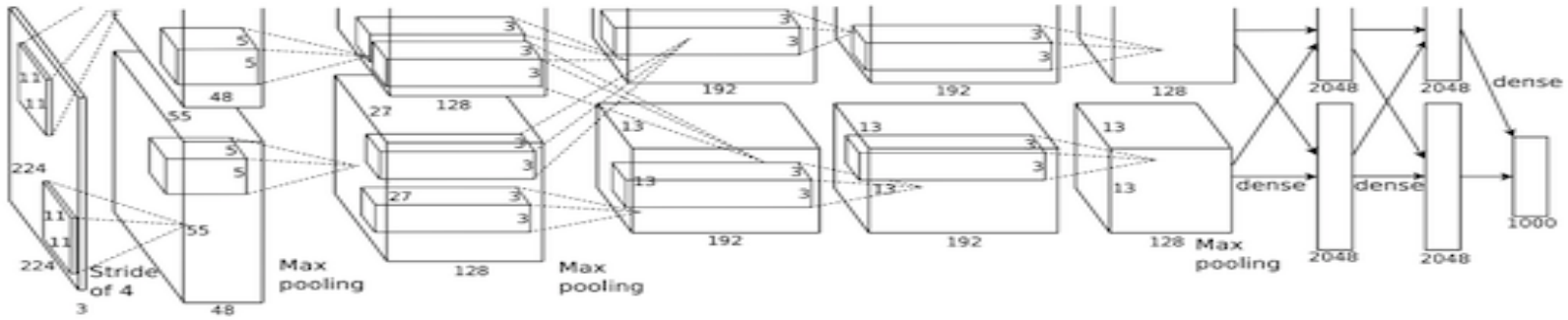
NAVER LABS Europe

First deep learning approaches

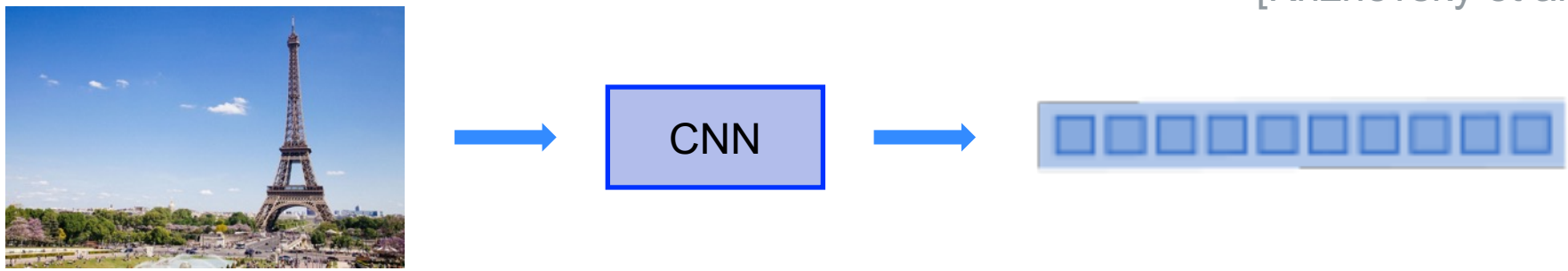
Pretrain network for classification and use it as a feature extractor

- Representations are compact and fast at test time!

AlexNet



[Krizhevsky et al. NIPS 2012]



Input image

NAVER LABS Europe

Limitations of the naïve deep approach

Obvious limitations:

- Network trained for generic classes – intra-class generalization
- Low resolution and distort aspect ratio

➢ Underwhelming results

Solution 1

- Combine with more standard approaches: hybrid methods

Solution 2

- Improve the pipeline to tailor it to the retrieval task: recent deep approaches

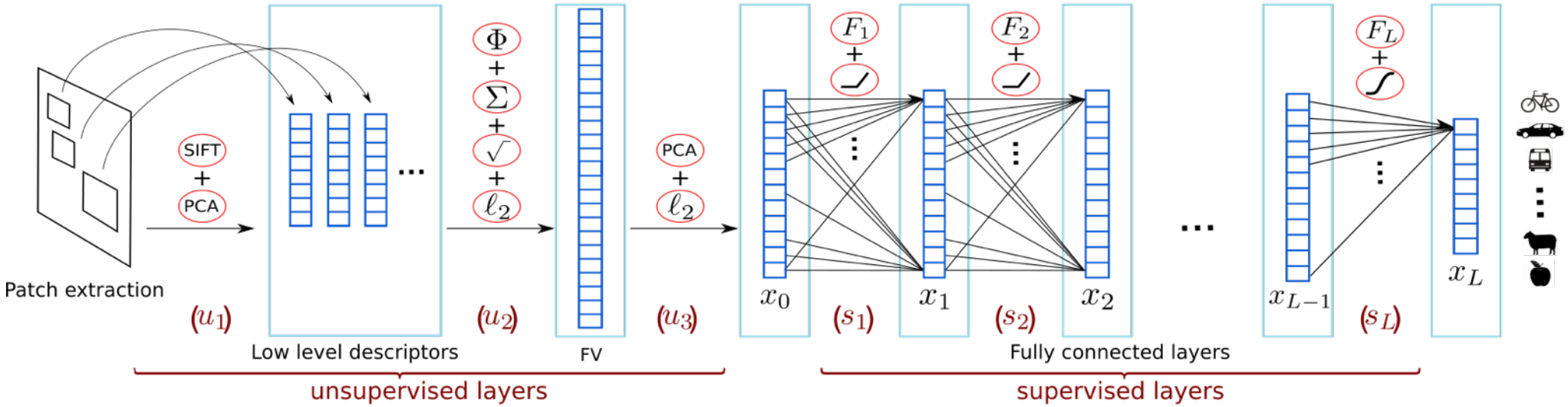
NAVER LABS Europe

Combining Fisher Vector and Fully connected layers

Principle:

[Gordo et al. CVPR12, Perronnin & Larlus. CVPR15]

- Fisher Vector representation combined with fully-connected layers



NAVER LABS Europe

Limitations of the naïve deep approach

Obvious limitations:

- Network trained for generic classes – intra-class generalization
- Low resolution and distort aspect ratio

➢ Underwhelming results

Solution 1

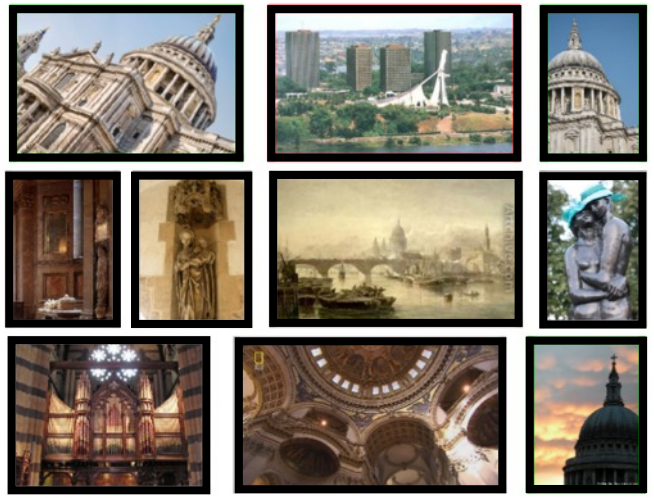
- Combine with more standard approaches: hybrid methods

Solution 2

- Improve the pipeline to tailor it to the retrieval task: recent deep approaches

NAVER LABS Europe

Training for the visual search



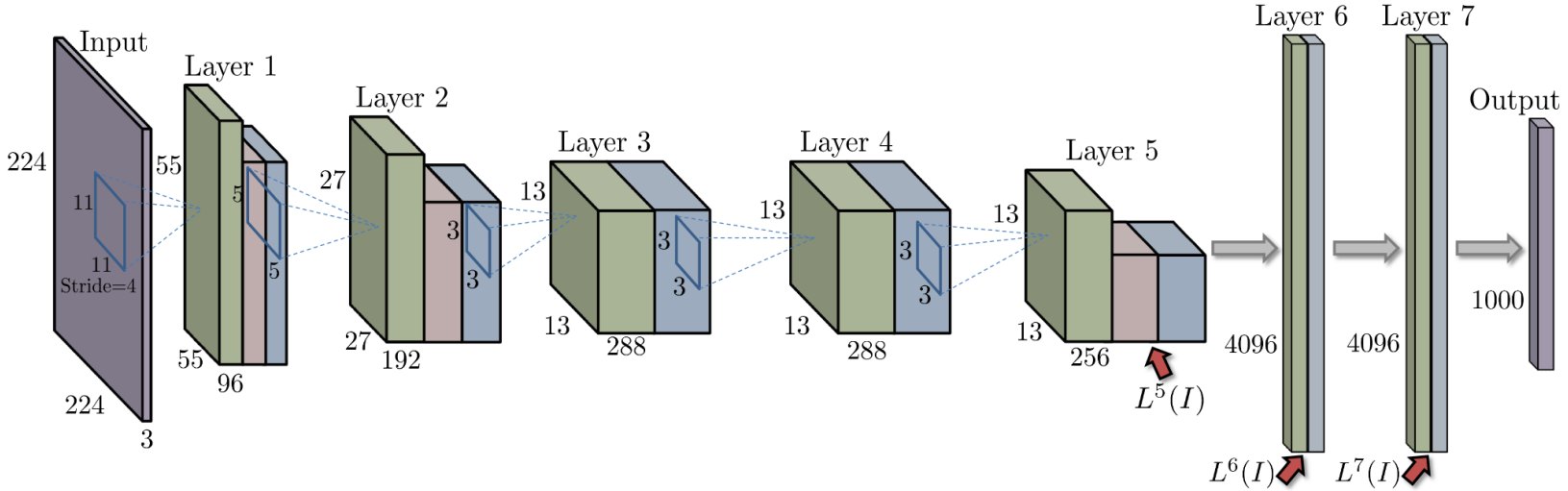
Collect and leverage an appropriate training set

- ~200K images
- ~600 different landmarks

Fine-tune the network

- Images resized to be 224x224
- Softmax classification loss

[Babenko et al, Neural codes @ ECCV14]



NAVER LABS Europe

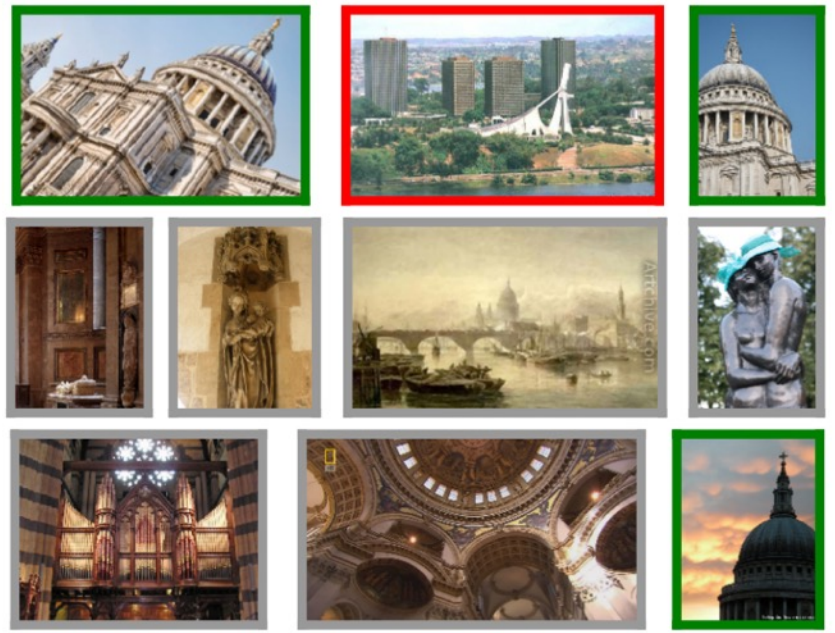
Training for the visual search

Limitations

[Babenko et al, Neural codes @ ECCV14]

- Again, images are distorted
- Fine-tune network on relevant images still with a classification loss

- Public landmark dataset is very noisy



NAVER LABS Europe

Training for the visual search

What can be improved?

1. Training data:

   Public landmark dataset is very noisy, needs to be cleaned automatically

2. Architecture:

   Small details are important for instance level retrieval: need to accommodate high resolution, undistorted images during training

3. Training objective:

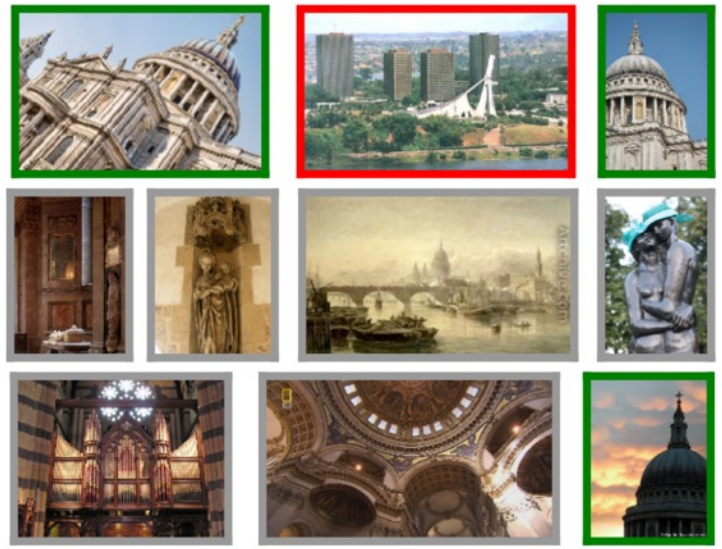
   We should train explicitly for retrieval, not for classification

1. Training Data

Public dataset of landmark images

- ~200K images
- ~600 different landmarks (Rome colosseum, Big Ben…)

[Gordo et al. ECCV 16, IJCV17]



NAVER LABS Europe

1. Training Data

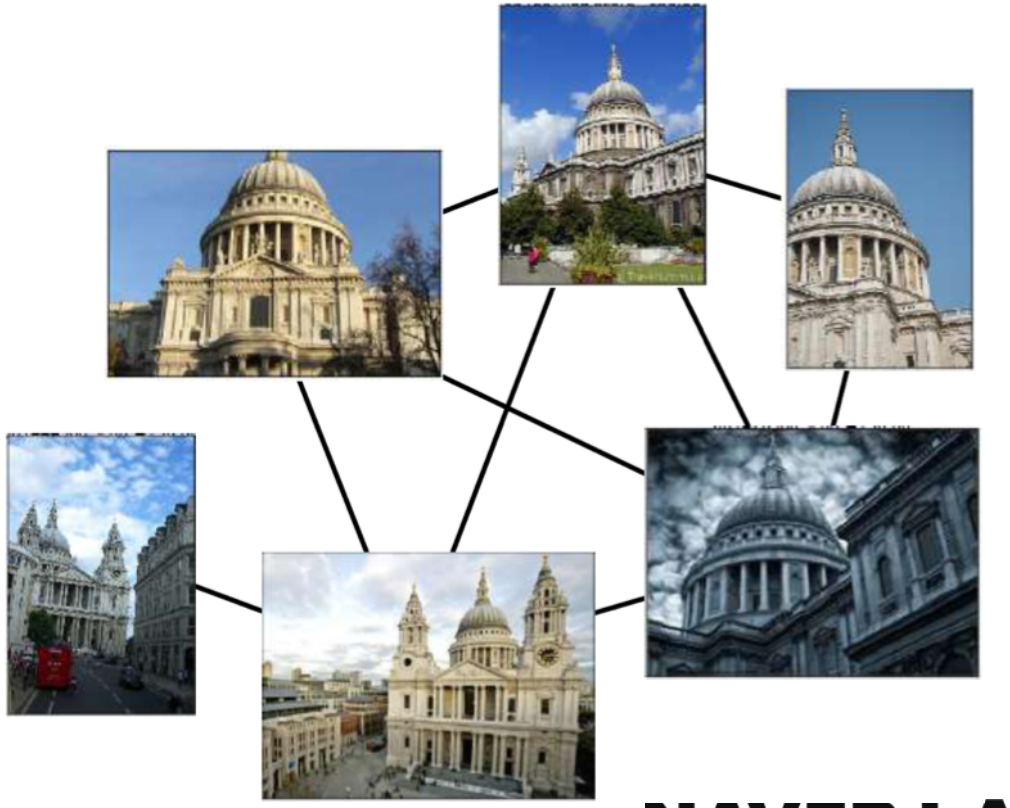
Public dataset of landmark images

- ~200K images
- ~600 different landmarks (Rome colosseum, Big Ben…)

[Gordo et al. ECCV 16, IJCV17]

Automatic cleaning of existing dataset

- Using standard method based on local descriptors to connect pairs of relevant images
  - Interest point detector: Hessian-Affine
  - Description of the local descriptors with SIFT
  - Pairwise matching between local descriptors
  - Geometrical verification with RANSAC
- Graph construction
  - Keep the largest connected component per class



NAVER LABS Europe

1. Training Data

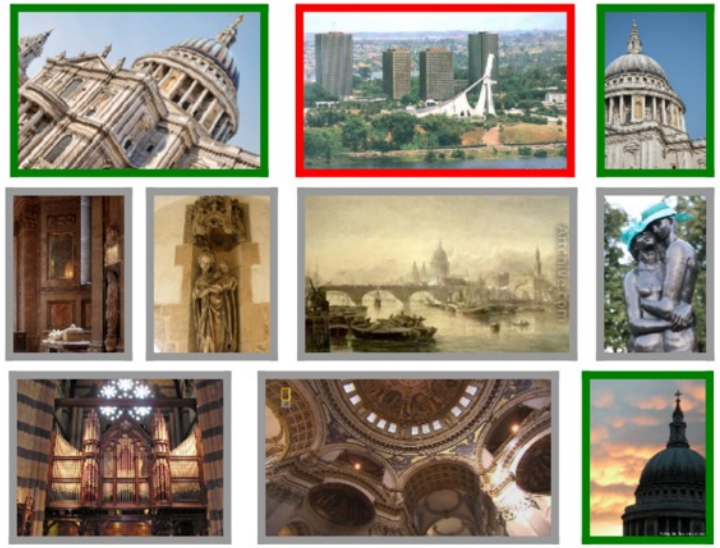
Public dataset of landmark images

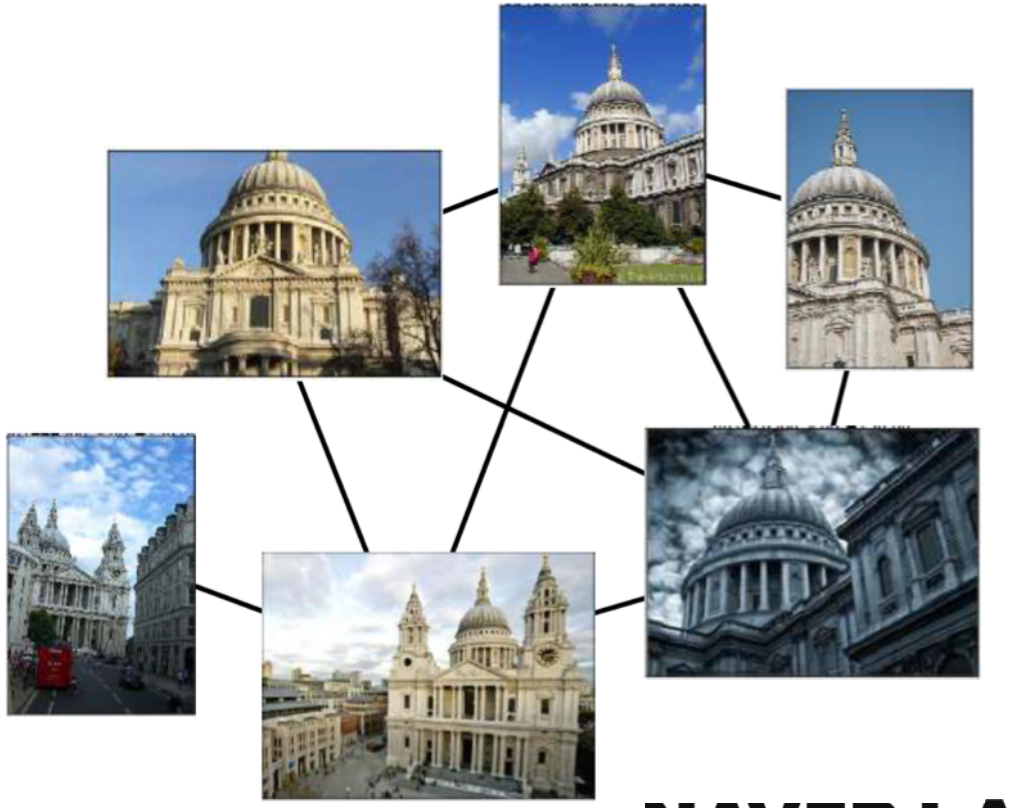
- ~200K images
- ~600 different landmarks (Rome colosseum, Big Ben…)

[Gordo et al. ECCV 16, IJCV17]

Automatic cleaning technique, resulting in:

- 40K spatially verified images
- Approximate bounding box annotations





NAVER LABS Europe

2. Architecture

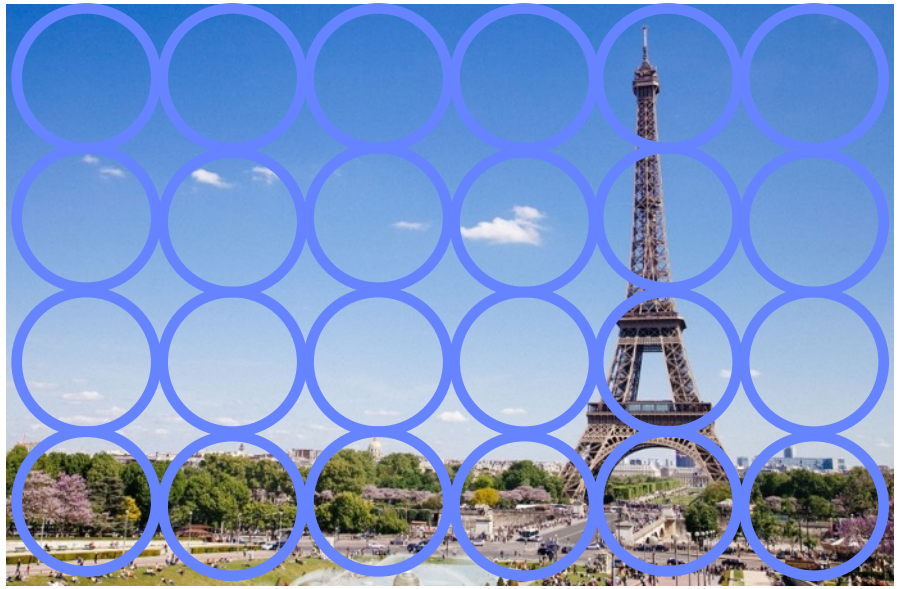
R-MAC descriptor

[Tolias et al, ICLR16]

Input image

CNN

Local features



NAVER LABS Europe

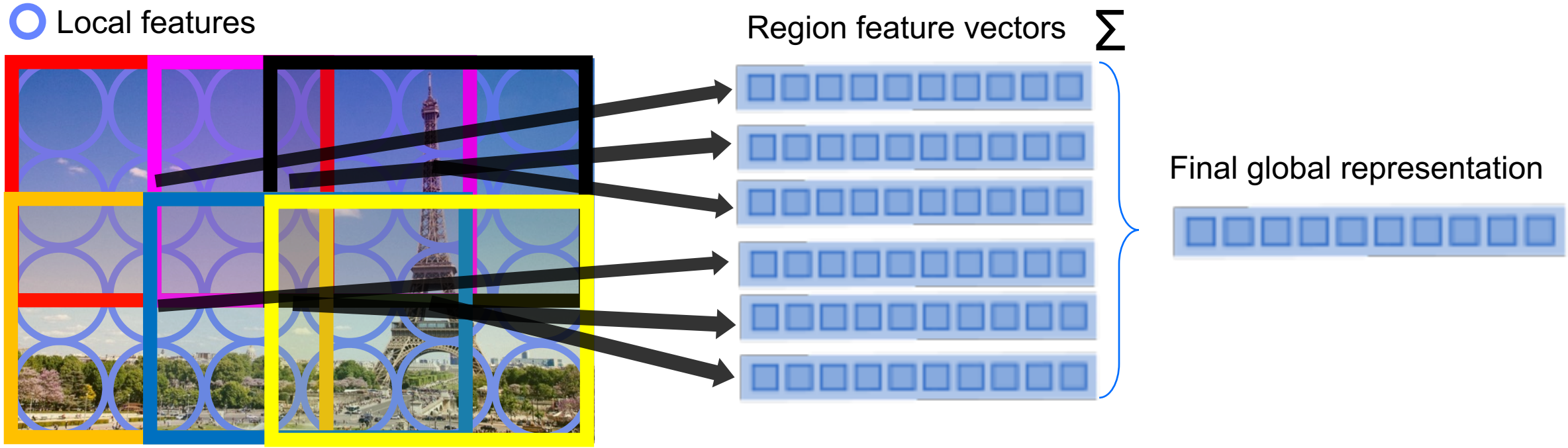
2. Architecture

R-MAC descriptor

[Tolias et al, ICLR16]

Advantages

- no aspect ratio distortion
- can encode high resolution images
- fast comparison with the dot product



NAVER LABS Europe

2. Architecture

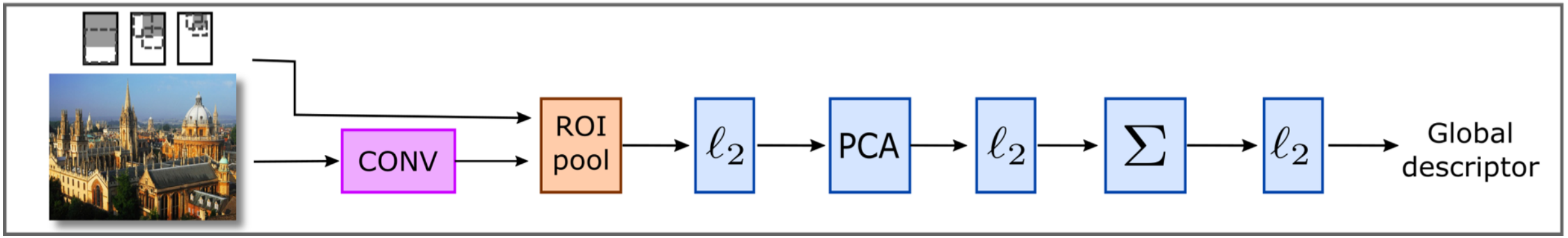
R-MAC descriptor

- CNN as a local feature extractor

Two key observations

[Gordo et al. ECCV 16, IJCV17]

1. The aggregation steps can be integrated inside the network:



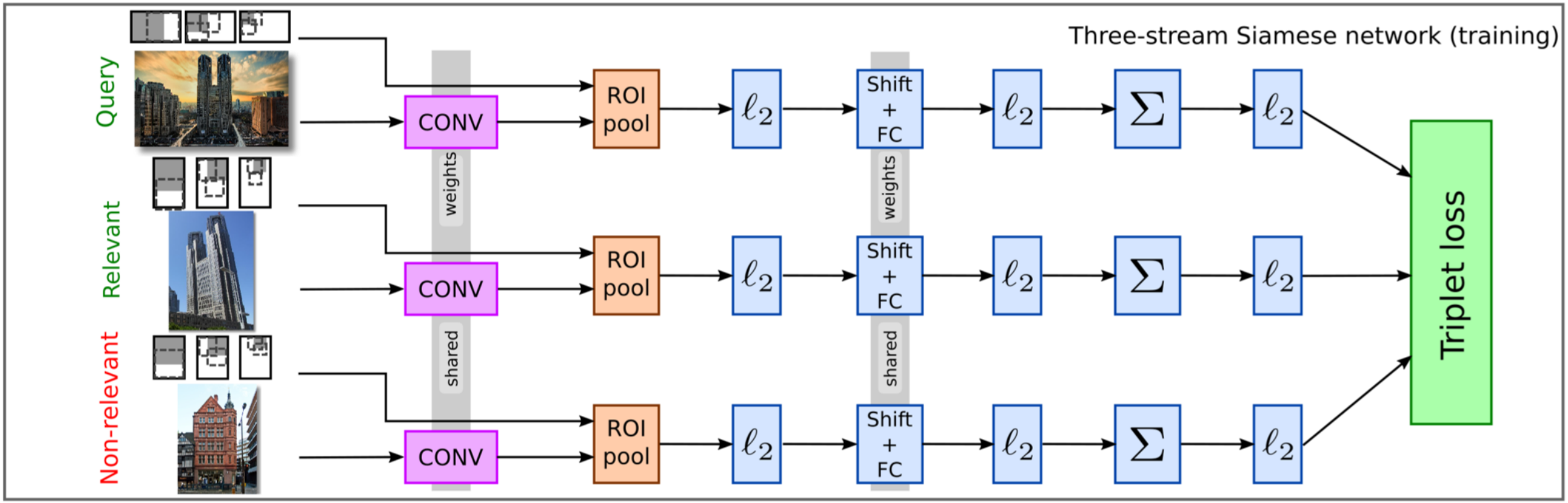
2. Every step is differentiable ⇒ **the model can be trained end-to-end!**

NAVER LABS Europe

3. Training for retrieval

Learning to rank

[Gordo et al. ECCV 16, IJCV17]
[Radenovic et al. ECCV 16, PAMI18]



NAVER LABS Europe

3. Training for retrieval

Triplet loss

[Gordo et al. ECCV 16, IJCV17]

$$L_v(q, d^+, d^-) = \frac{1}{2}\max(0, m - \phi_q^T \phi_+ + \phi_q^T \phi_-)$$

Query

Relevant

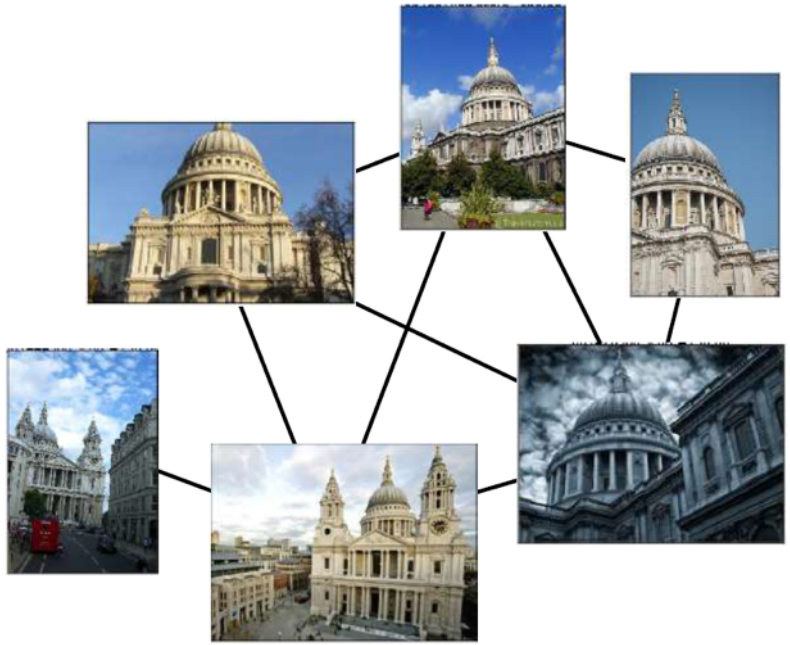
Non-relevant

NAVER LABS Europe

Summary

Problems:

1. Features from ImageNet are not good at intra-class discrimination



Solutions:

1. Train on an **automatically cleaned** dataset of landmarks

NAVER LABS Europe

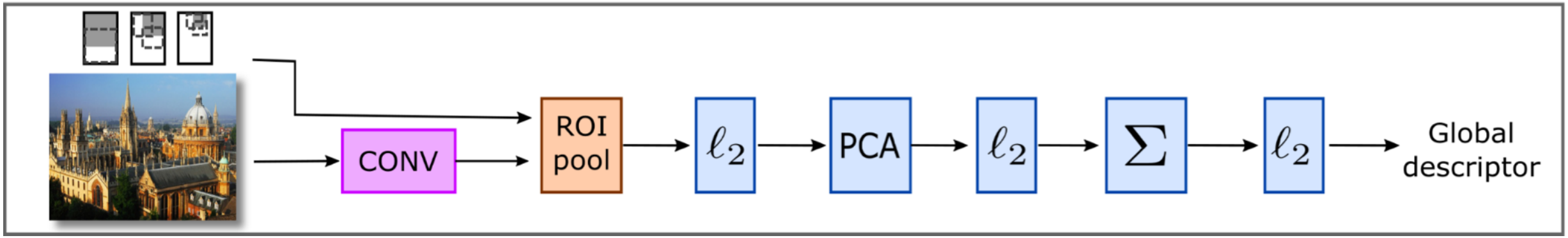
Summary

Problems:

1. Features from ImageNet are not good at intra-class discrimination
2. Usual architectures work on small crops of the image and at low resolution

Solutions:

1. Train on an **automatically cleaned** dataset of landmarks
2. Use an **architecture that preserves image details**



NAVER LABS Europe

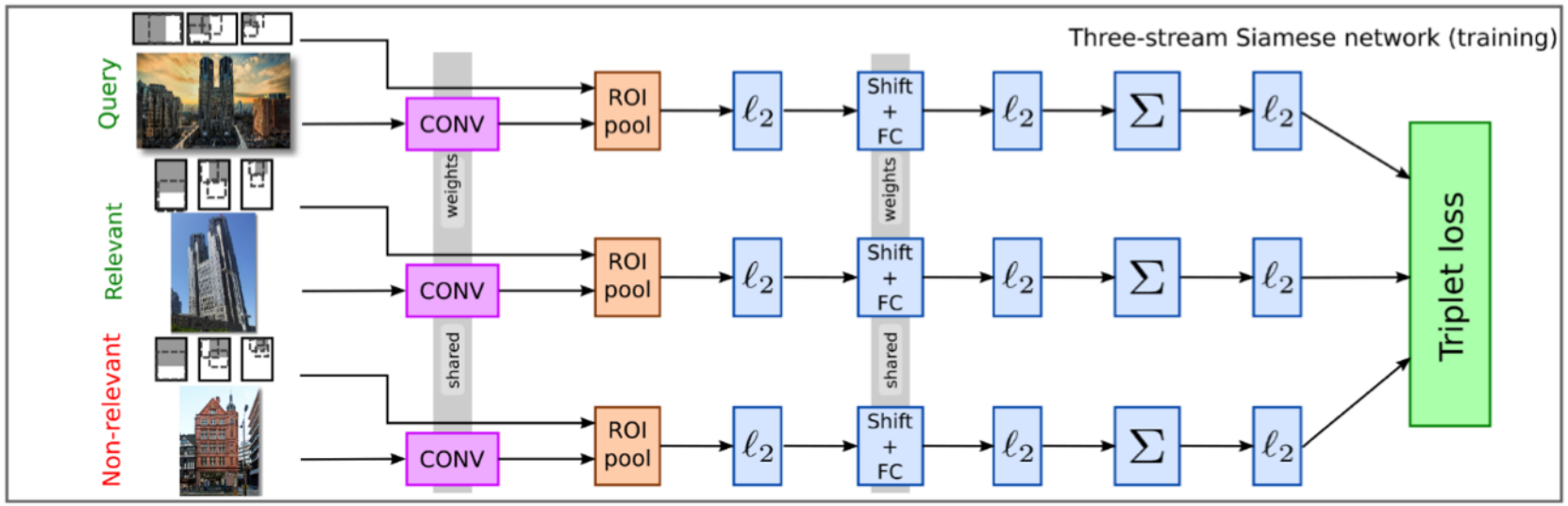
Summary

Problems:

1. Features from ImageNet are not good at intra-class discrimination
2. Usual architectures work on small crops of the image and at low resolution
3. Networks are typically trained for classification

Solutions:

1. Train on an **automatically cleaned** dataset of landmarks
2. Use an **architecture that preserves image details**
3. Train the network with a **ranking loss**



NAVER LABS Europe

Qualitative results

[Gordo et al. ECCV 16]

query

top retrieved

NAVER LABS Europe

Impact of fine-tuning an ImageNet model

Where do conv5 neurons fire? (initial VGG model) [Gordo et al. ECCV 16]

Before

After

NAVER LABS Europe

Impact of fine-tuning an ImageNet model

Where do conv5 neurons fire? (initial VGG model) [Gordo et al. ECCV 16]

Before

After

NAVER LABS Europe

Other aggregation techniques

SPoC

[Babenko and Lempitsky. ICCV15]

- Sum-pooling aggregation over convolutional features
- Center prior for spatial weighting and uniform channel weighting

CroW

[Kalantidis et al. W@ECCV16]

- Non-parametric scheme for spatial- and channel-wise weighting before sum-pooling aggregation
- Boosts the effect of highly activate spatial responses and regulates burstiness

Generalized-Mean pooling

[Radenovic et al. PAMI18]

- Trainable Generalized-Mean pooling layer that generalizes max and average pooling
- **Top performing approach**

NAVER LABS Europe

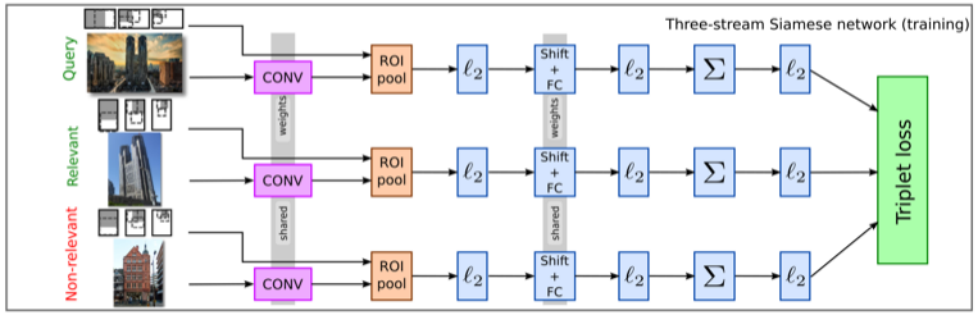
Summary: different losses

- Contrastive loss (pairwise):

  F. Radenovic, G. Tolias, and O. Chum. *CNN image retrieval learns from BoW: Unsupervised fine-tuning with hard examples.* ECCV 2016 & PAMI 2018

- Triplet loss (pairwise):

  A. Gordo, J. Almazan, J. Revaud, and D. Larlus. *End-to-end learning of deep visual representations for image retrieval.* ECCV 2016 & IJCV 2017



Next: more sophisticated losses

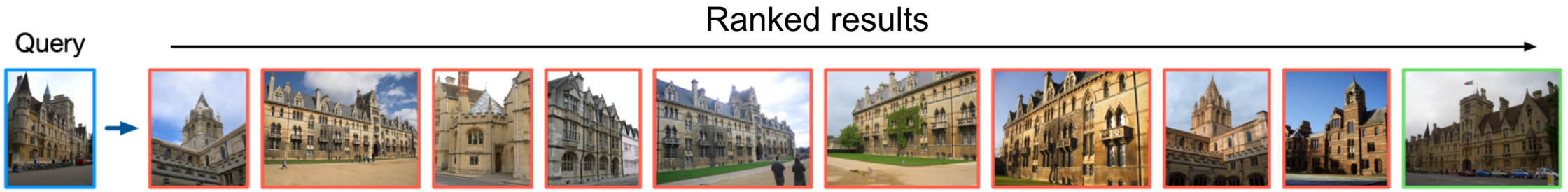
NAVER LABS Europe

Back to the motivation

- Query by example: results should be « well ranked »
- How should measure this?

Back to the motivation

- Query by example: results should be « well ranked »
- How should measure this?

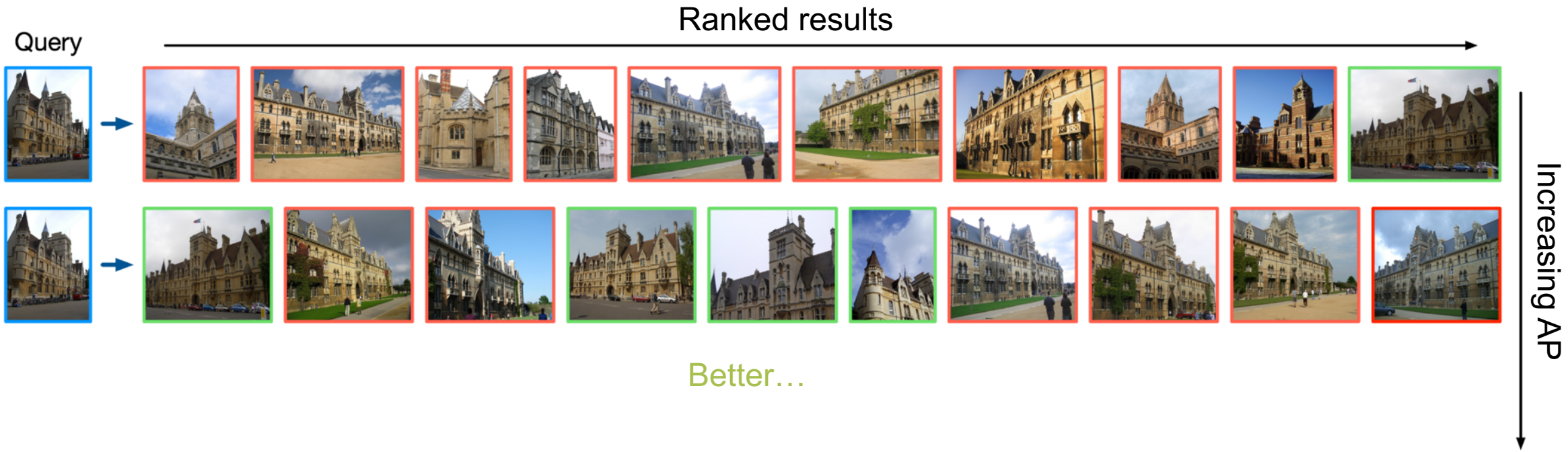


Bad!

NAVER LABS Europe

Back to the motivation

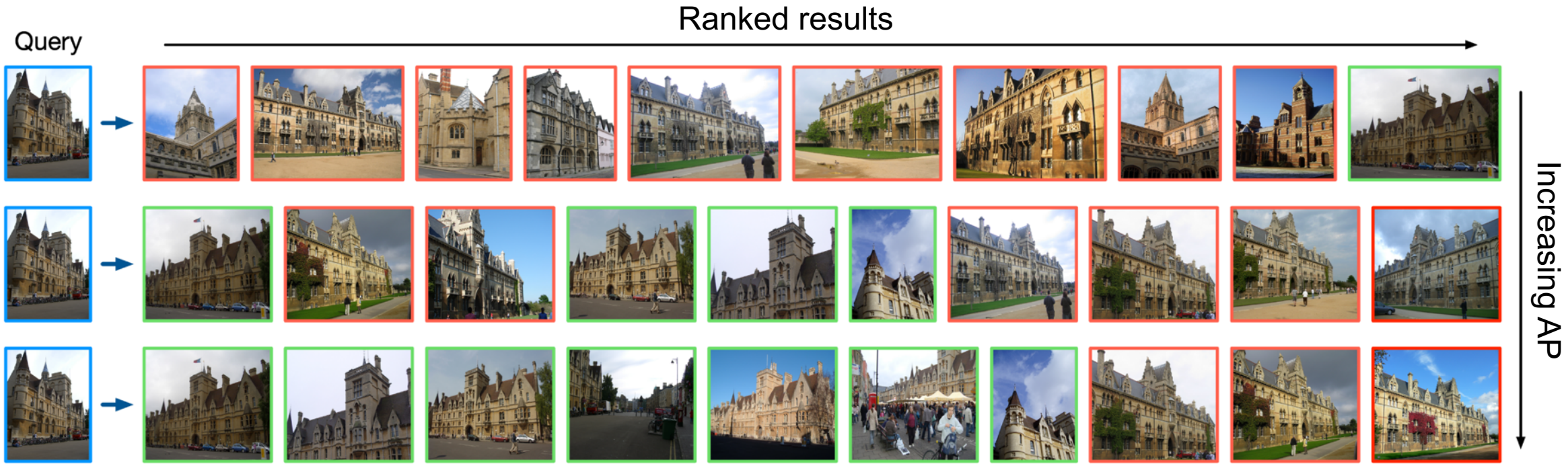
- Query by example: results should be « well ranked »
- How should measure this?



NAVER LABS Europe

Back to the motivation

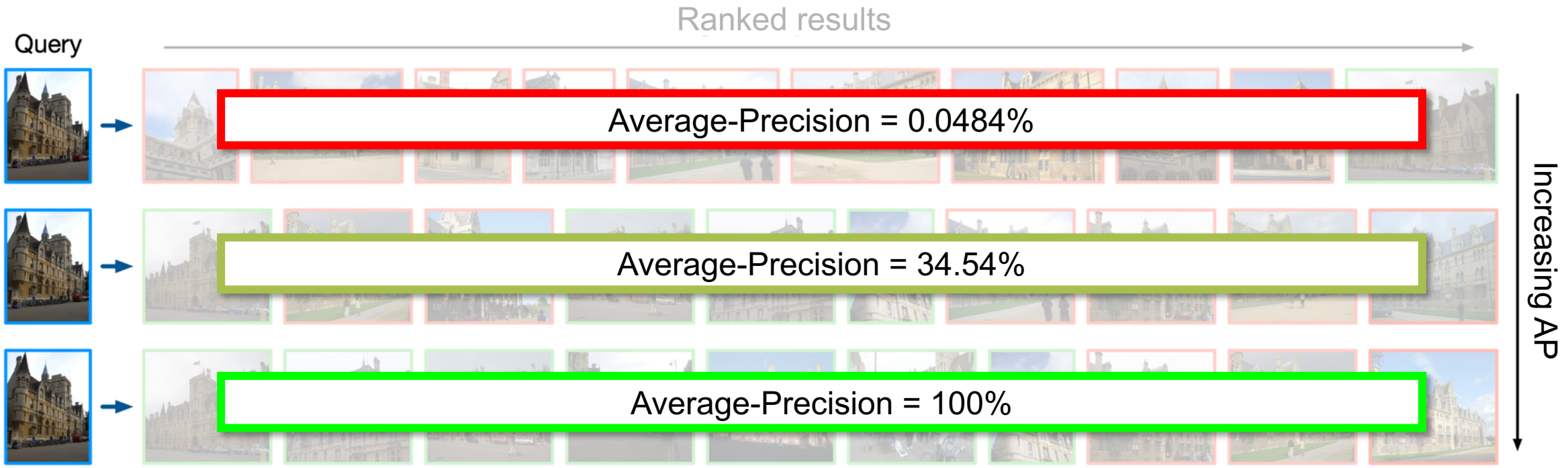
- Query by example: results should be « well ranked »
- How should measure this?



NAVER LABS Europe

Back to the motivation

- Query by example: results should be « well ranked »
- How should measure this?



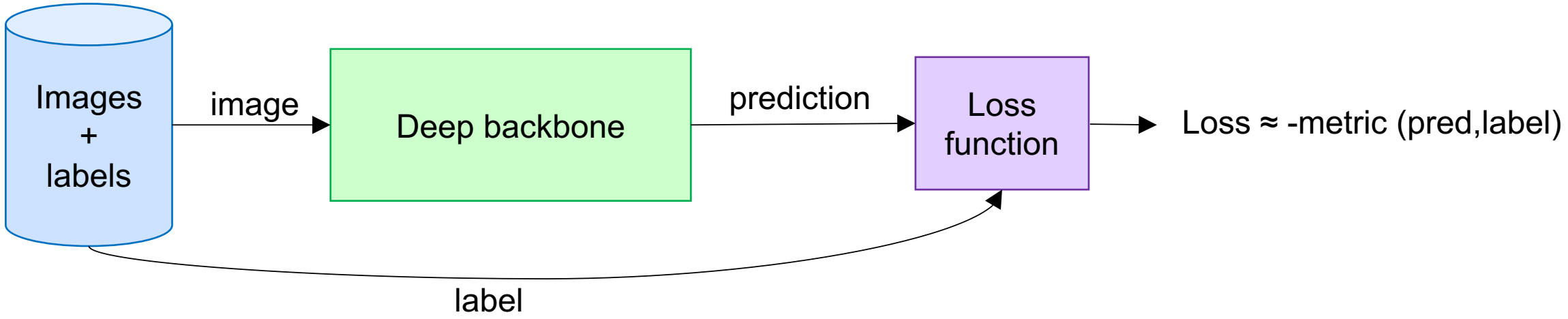
NAVER LABS Europe

What is Average-Precision?

- Measure how query results are « well ranked »
- Average precision: score in [0,1] (often expressed in percents)

Successful deep learning recipe

1. Find a large-scale annotated dataset for your task
2. Take your favorite deep backbone (ResNet, GoogLeNet, DenseNet,..)
3. Take your evaluation metric ➔ express it as a *loss*



NAVER LABS
Europe

What about Average-Precision?

- Measure how query results are « well ranked »
- Average precision: score in [0,1] (often expressed in percents)

Then why not train for Average Precision (AP)?

- AP is non-differentiable because it involves sorting elements

➔ Cannot be used as a loss

The learn-to-rank framework

- There exists surrogate loss functions for ranking:
  - Pointwise losses: e.g. regression, margin loss
  - Pairwise losses: e.g. contrastive loss, **triplet loss**

  → local losses
  - **Listwise losses**

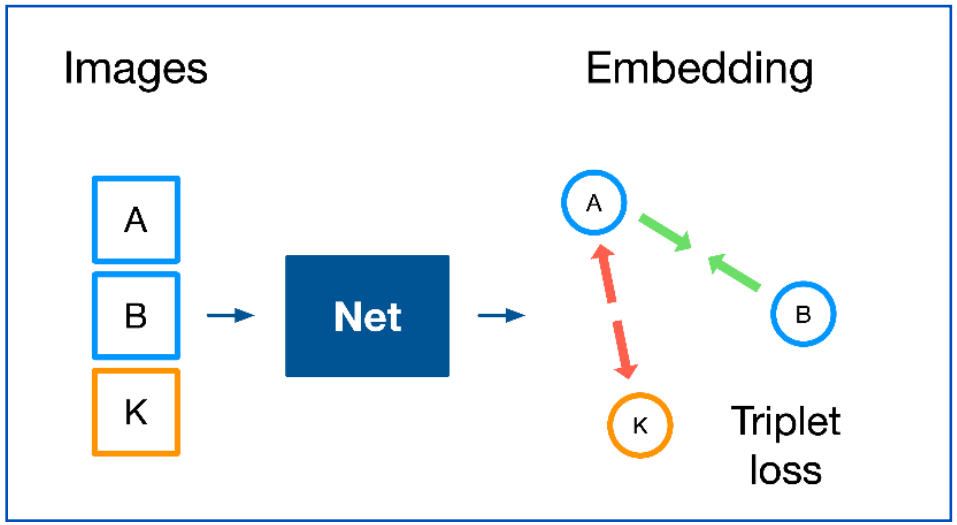
  → global losses

Learning with a local loss

[Schroff@CVPR15]

Example: the triplet loss

- *Take 3 images A, B, K (one query, one relevant, one non-relevant)*
- *Compute 3 embedings*
- *Pull closer A and B*
- *Push apart A and K*
- Repeat many times



Using triplets:
3 images at a time

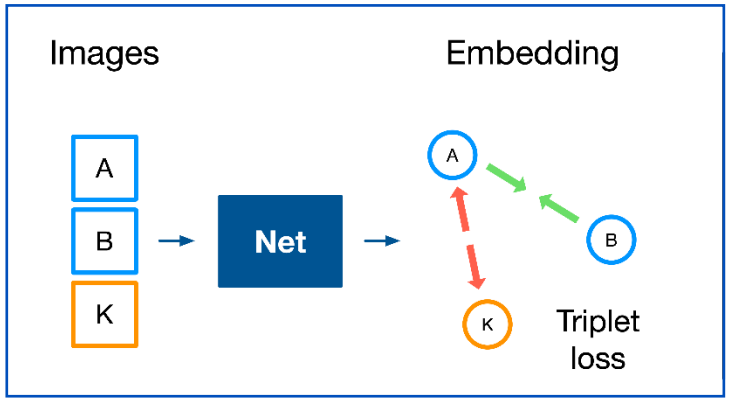
Limitations

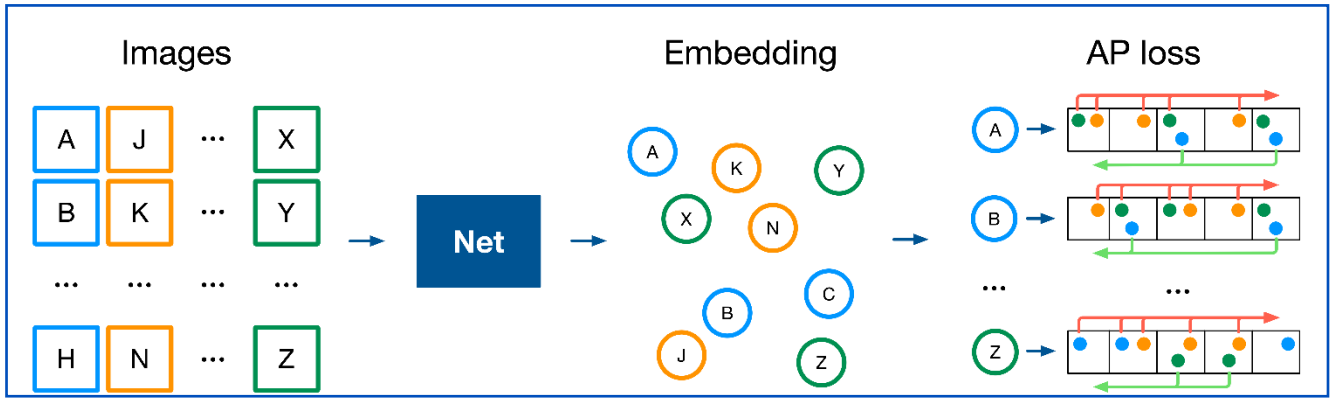
- Requires many iterations
- Triplets need to be sampled with care

[Hardwood@ICCV17, Manmatha@ICCV17, Mishchuk@NeurIPS17, *Gordo@IJCV17, Hermans@ArXiv17, Faghri@BMVC18, etc.*]

NAVER LABS Europe

Training for Average-Precision





Using triplets:
3 images at a time

Using a listwise loss:
Directly optimize the AP on
thousands of images simultaneously!

NAVER LABS Europe

Training for Average-Precision

It is in fact possible to have a differentiable *version of AP*

Key idea: implementing « *sorting* » differently

Most approaches are based on divide-and-conquer

➔ non-differentiable

But you can use histogram binning! [Revaud@ICCV19]

➔ well differentiable if you use a smooth density estimator (kernel)

Evaluation on standard benchmarks

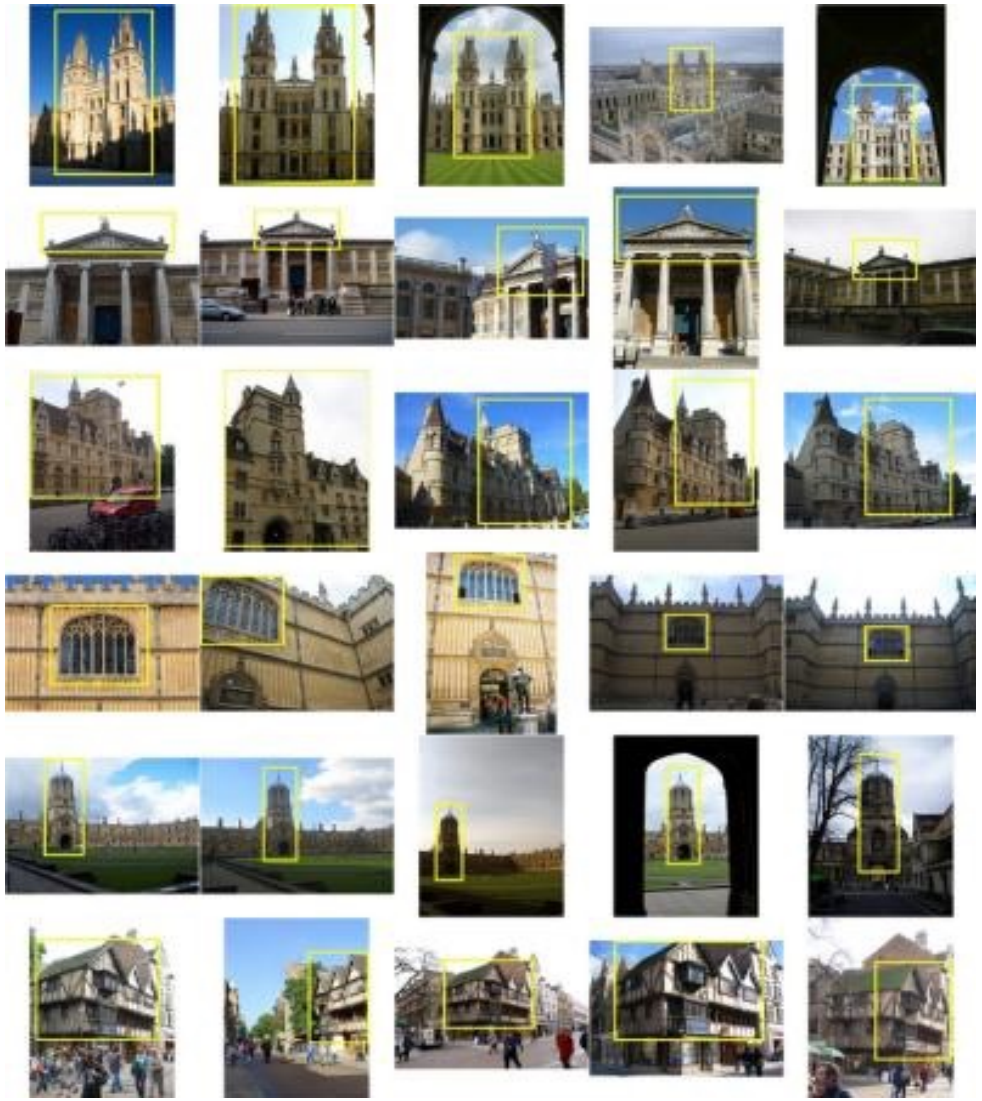
Oxford dataset

- 5,000 images
- 55 queries
- 11 landmarks

Evaluation

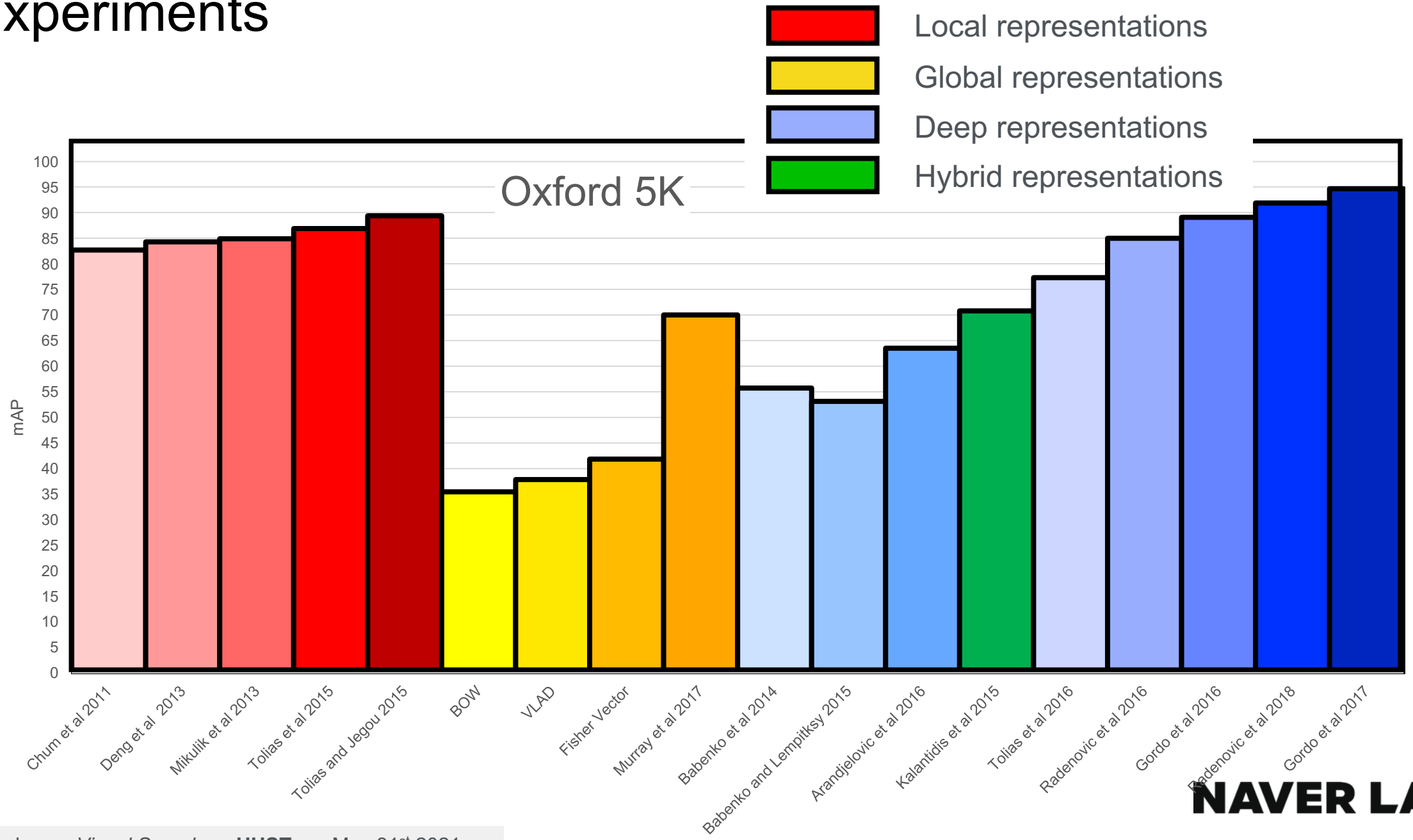
- mean Average Precision (mAP)

[Philbin et al. CVPR07]



NAVER LABS Europe

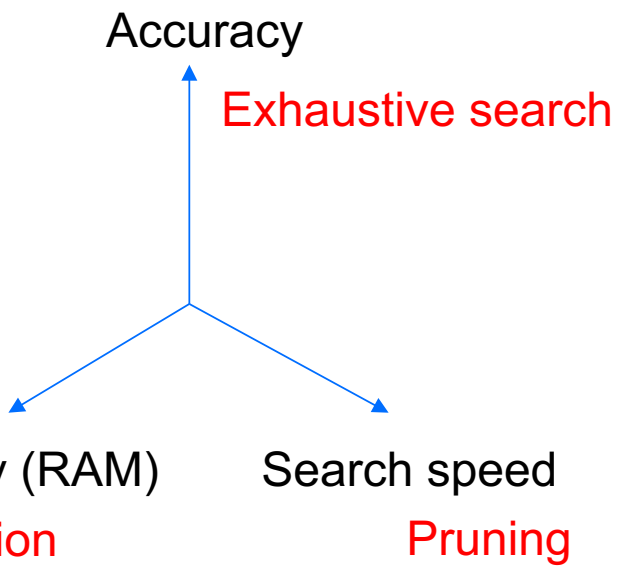
Experiments



NAVER LABS Europe

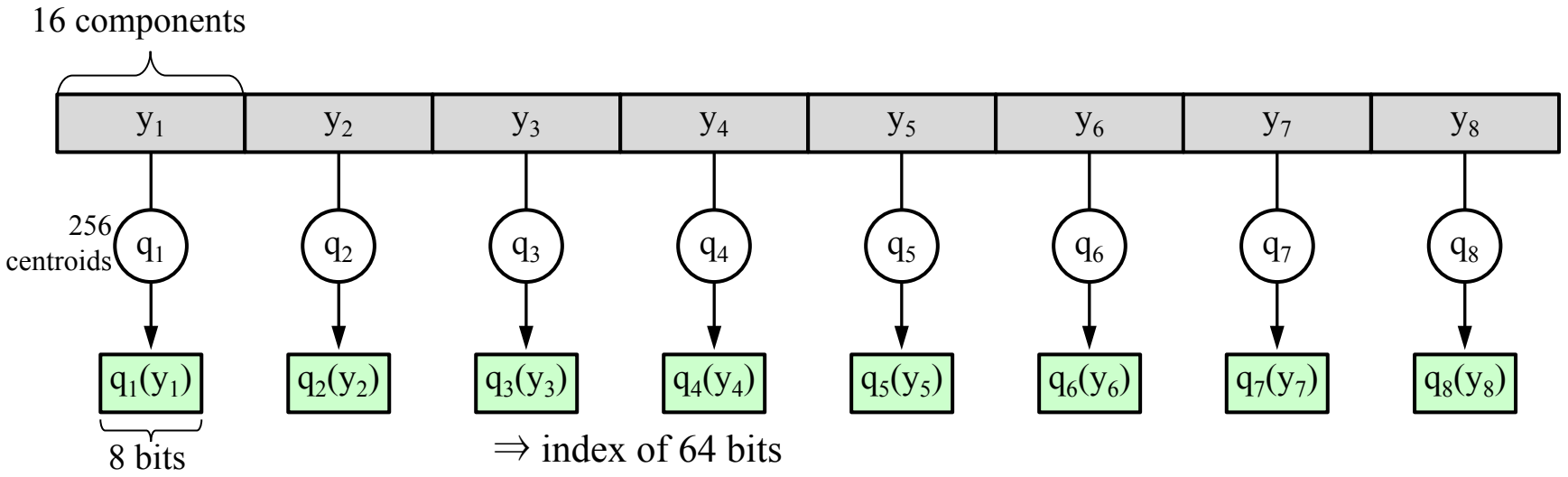
Ingredients of a full system 1/2

Trade-offs in similarity search



Compression

- Product Quantization (PQ)



[see Hervé Jegou's tutorials]

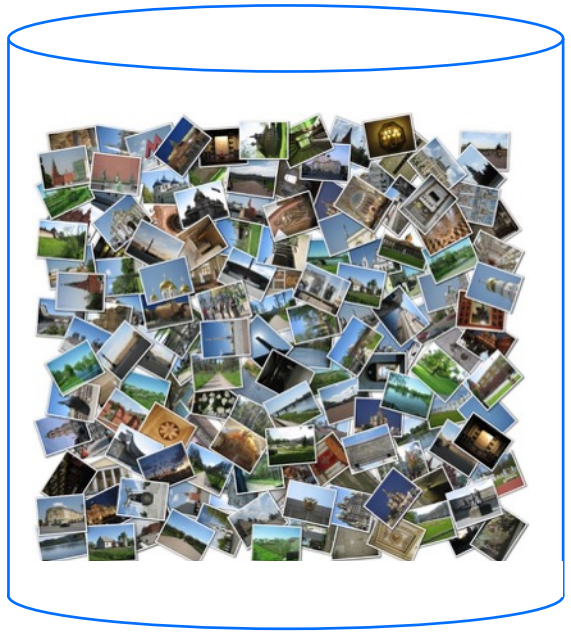
NAVER LABS Europe

Ingredients of a full system 2/2

Leveraging the topology of the gallery set

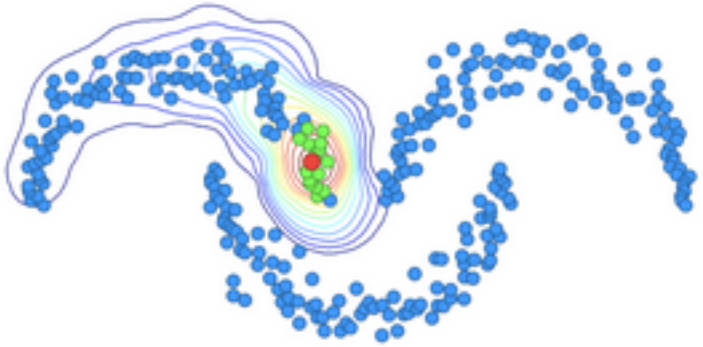
- Query expansion
  - Process the list of results (e.g. spatially verify)
  - If some images are good, use them to process some other augmented queries

[Chum et al. ICCV07, Chum et al CVPR11]



- Diffusion

[Iscen et al. CVPR17, CVPR18]



NAVER LABS Europe

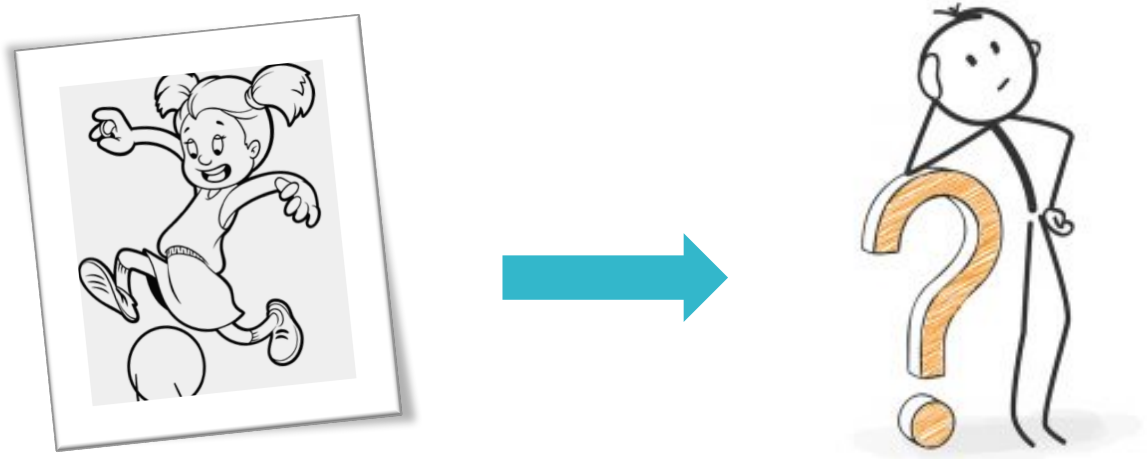
Outline

Object Search

Semantic Retrieval

NAVER LABS
Europe

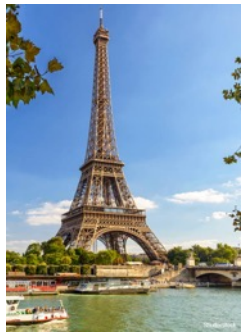
Semantic Retrieval

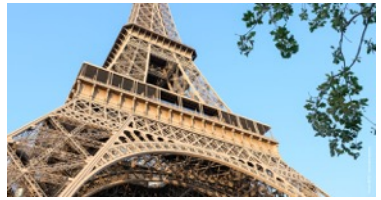


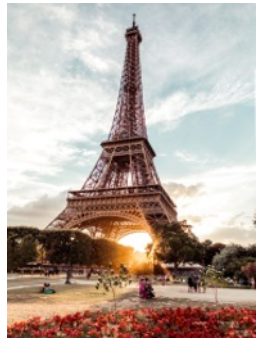
NAVER LABS
Europe

Standard Image Retrieval

Query: one example view of an **instance**







[Landmark Retrieval Challenge – CVPR18, CVPR19, ECCV20]

Instance level Retrieval

NAVER LABS Europe

Beyond Standard Image Retrieval

Complex queries that involve the **full scene**

[*Gordo & Larlus. CVPR17*]

Semantic Image Retrieval

NAVER LABS Europe

Semantic image retrieval

Is the semantic retrieval task well-defined?

Human annotations

%?

%?

NAVER LABS Europe

Semantic image retrieval

Is the semantic retrieval task well-defined?

Human annotations

User study:

- 35 annotators
- 3,000 image triplets

Evaluation:

- Leave-one-user-out agreement score: 89.1 +- 4.6

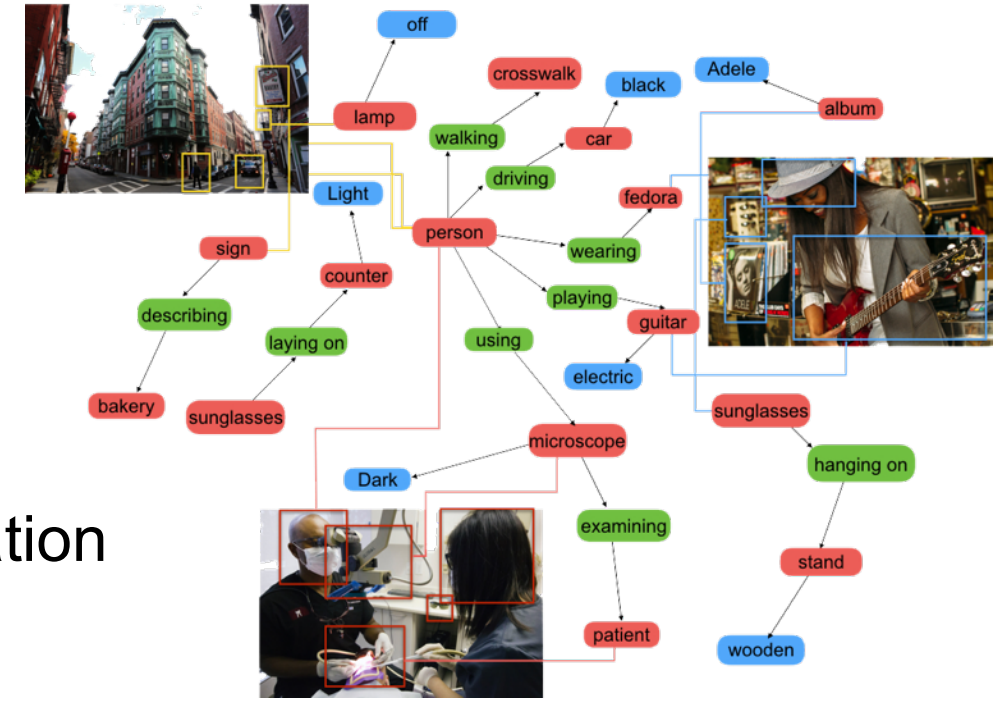
Semantic image retrieval

What could capture semantic information?

The **Visual Genome** dataset

- 108,077 Images
- 5.4 Million Region Descriptions
- 3.8 Million Object Instances

[Krishna@arxiv16]



Observation:

Textual description best capture semantic information

NAVER LABS Europe

Semantic image retrieval

Can we learn a compact visual representation for the semantic retrieval task?

Training data

- Assume **human captions** are available at training time: **privileged information**
- Leverage human captions as a **proxy** for semantic similarity

Semantic image retrieval

Human-generated captions

The woman in purple on the steps
the person has on boots
open purple umbrella
the person has on boots
Woman in a garden holding an umbrella

Semantically similar

a woman under a umbrella
brown leather boots on legs
black umbrella is open
step leading to a door

NAVER LABS Europe

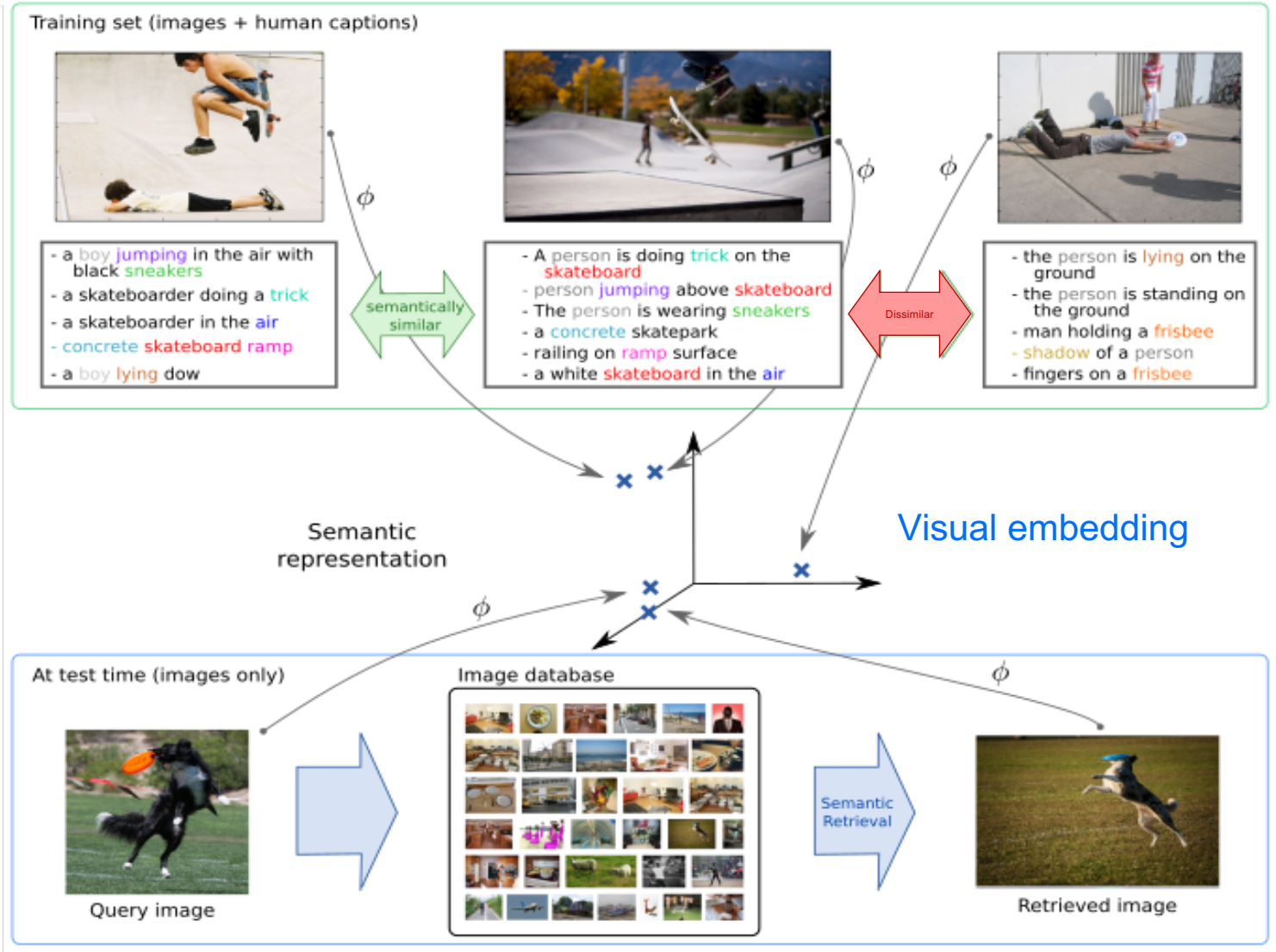
Semantic image retrieval

Building visual representations:

Intuition

For images with similar captions:

Visual representation close in the **semantic embedding space**



Learning a semantic embedding

Learning to rank = triplet loss

$$L_v(q, d^+, d^-) = \frac{1}{2}\max(0, m - \phi_q^T\phi_+ + \phi_q^T\phi_-)$$

[Gordo et al. ECCV 16, IJCV17]

NAVER LABS Europe

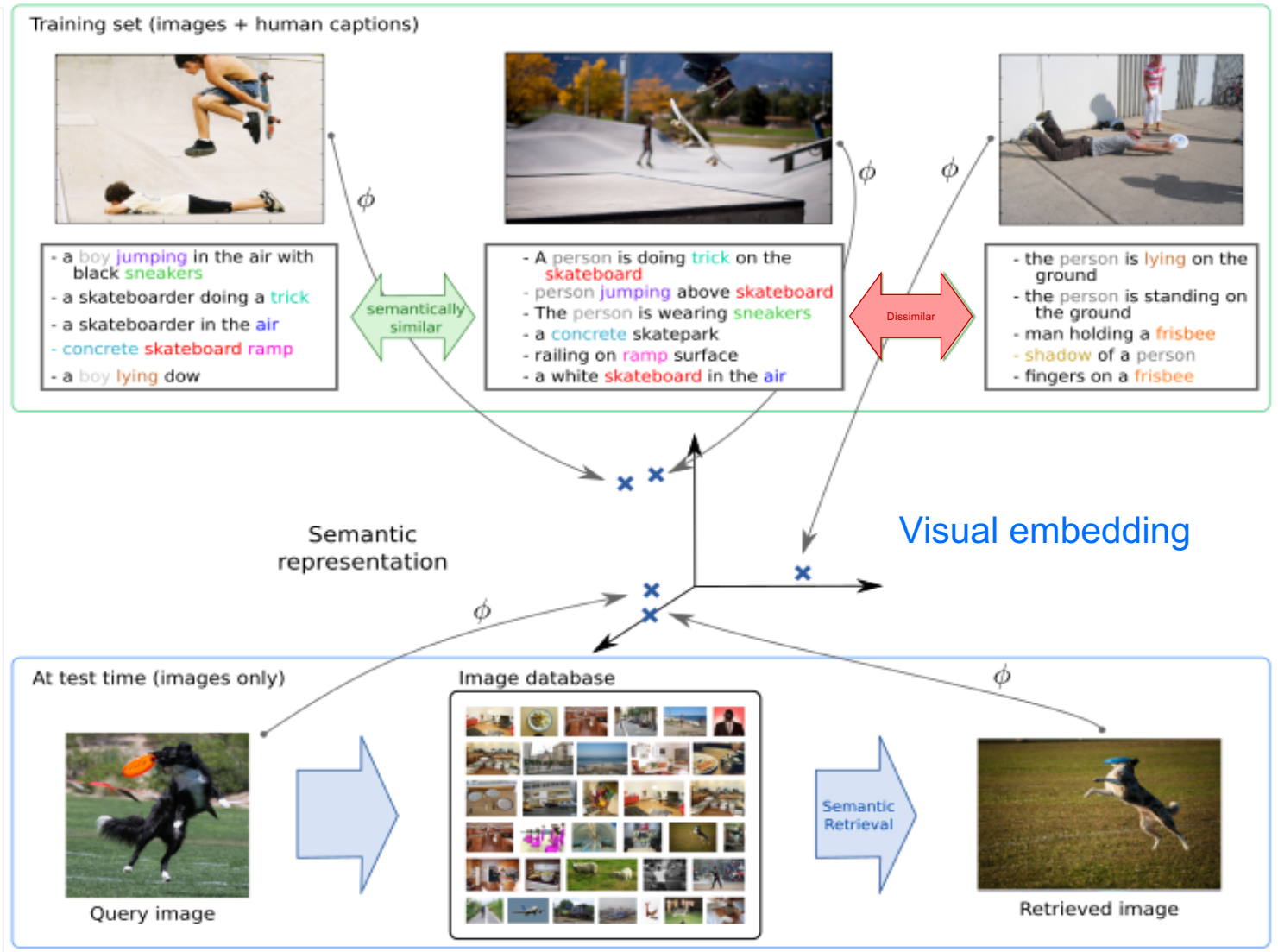
Semantic image retrieval

Building visual representations:

Intuition

For images with similar captions:

Visual representation close in the **semantic embedding space**

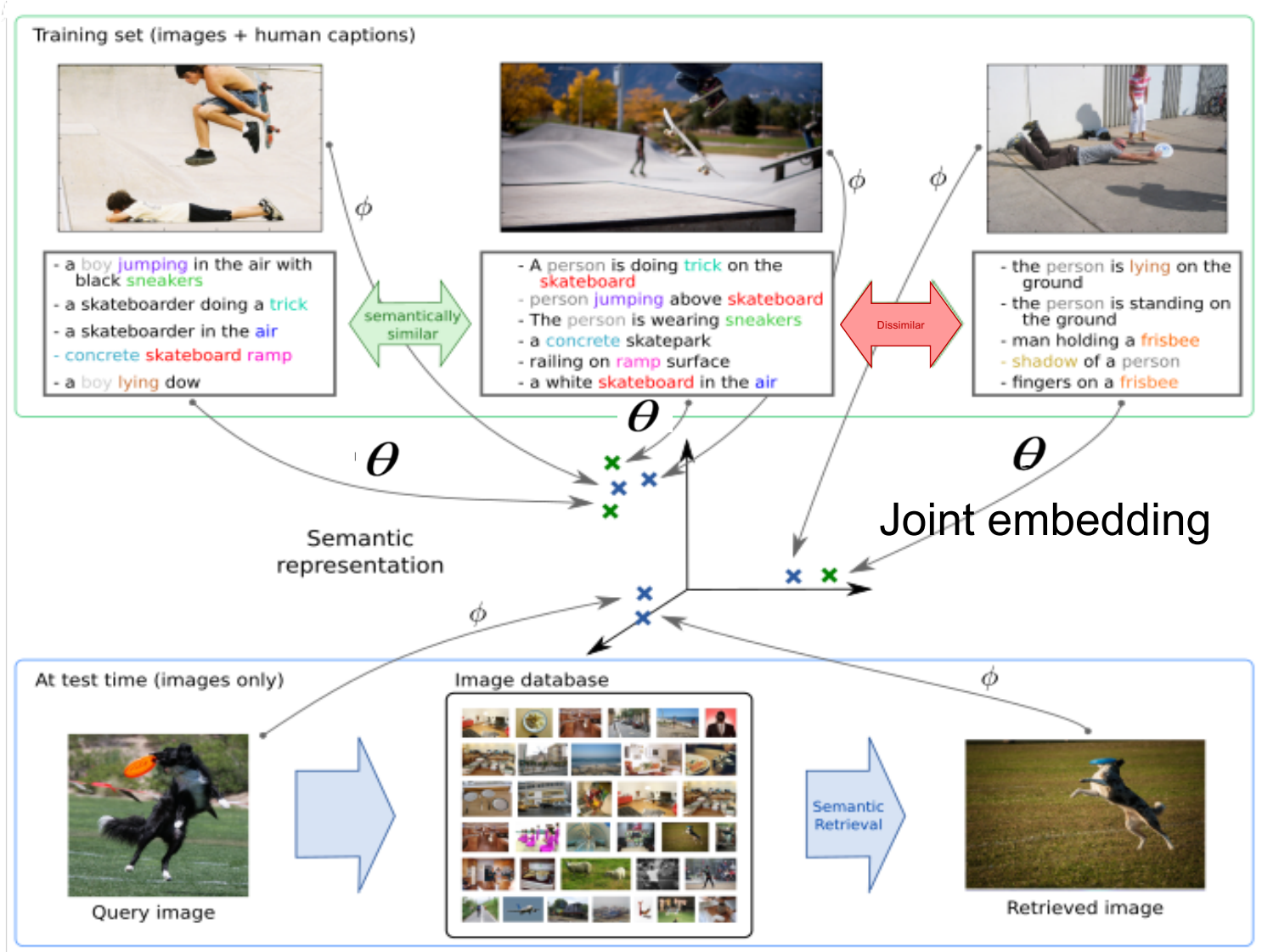


Semantic image retrieval

Training the model

The visual representation of images with similar captions are close in the **semantic embedding space**

The textual representation of corresponding captions are close in the **semantic embedding space**



Learning a semantic embedding

Visual loss:

$$L_v(q, d^+, d^-) = \frac{1}{2} \max(0, m - \phi_q^T \phi_+ + \phi_q^T \phi_-)$$

Textual losses:

$$L_{t1}(q, d^+, d^-) = \frac{1}{2} \max(0, m - \phi_q^T \theta_+ + \phi_q^T \theta_-)$$

$$L_{t2}(q, d^+, d^-) = \frac{1}{2} \max(0, m - \theta_q^T \phi_+ + \theta_q^T \phi_-)$$

NAVER LABS Europe

Semantic retrieval: experiments

Dataset: the Visual Genome

- Training set: 80k images for training - captions are leveraged
- Validation set: 10k images - no caption
- Test set: 10k images - no caption

[Krishna@arxiv16]

We evaluate on the triplets: User-based score (US)

- **Agreement score** between visual predictions and **users**

Semantic retrieval: quantitative results

Evaluation measures

- **US**: agreement between model predictions and **users**

[Weston, WSABIE @IJCAI11]

	US
Text oracle	
Caption Tf-idf	76.3
Query by image	
Random (x5)	50.0 ± 0.8
Visual baseline (, V)	64.0
WSABIE (V+T, V)	67.8
Proposed (V, V)	76.9
Proposed (V+T, V)	77.2

[*Gordo & Larlus. CVPR17*]

NAVER LABS Europe

Semantic retrieval: qualitative results

Query

Top semantically retrieved

Baseline

After training

[*Gordo & Larlus. CVPR17*]

Semantic retrieval: qualitative results

The joint embedding allows for multimodal queries

[*Gordo & Larlus. CVPR17*]

NAVER LABS Europe

Semantically matching images: providing a visual explanation

[*Gordo & Larlus. CVPR17*]

NAVER LABS Europe

Hanoi University of Science and Technology

Thanks!