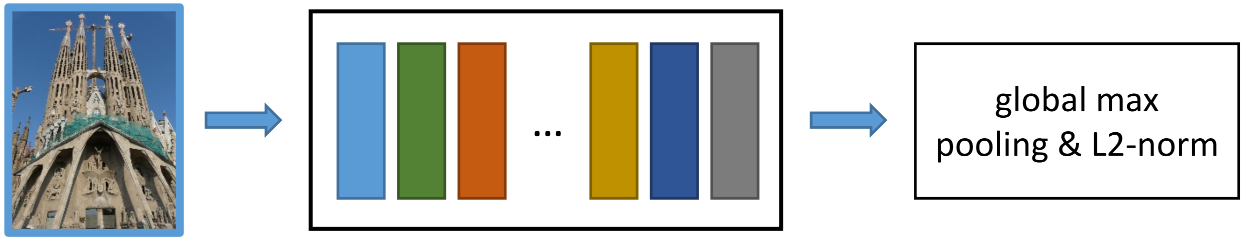


CNN Image Retrieval Learns from BoW: Unsupervised Fine-Tuning with Hard Examples

Filip Radenović Giorgos Tolias Ondřej Chum

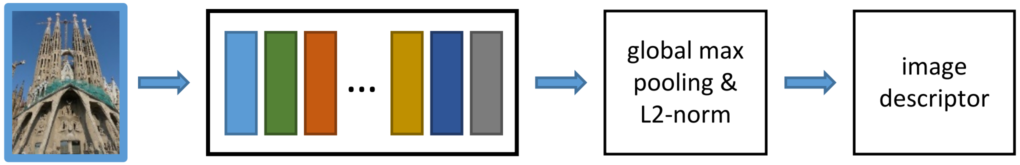
Center for Machine Perception, CTU in Prague

ECCV 2016

CNN Image Retrieval Learns from BoW: Unsupervised Fine-Tuning with Hard Examples

CNN Image Retrieval

- compact image descriptors
- Nearest Neighbor search

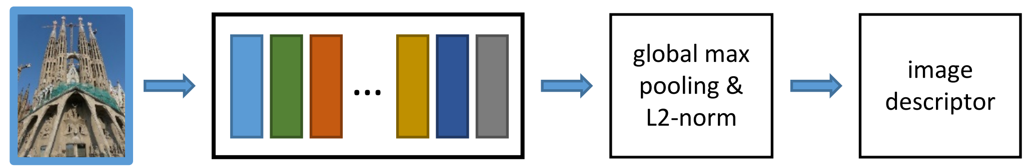


CNN Image Retrieval Learns from BoW: Unsupervised Fine-Tuning with Hard Examples

CNN Image Retrieval

compact image descriptors

Nearest Neighbor search



CNN Learning (Fine-Tuning)

start with CNN trained for different but similar task (reasonable parameters)

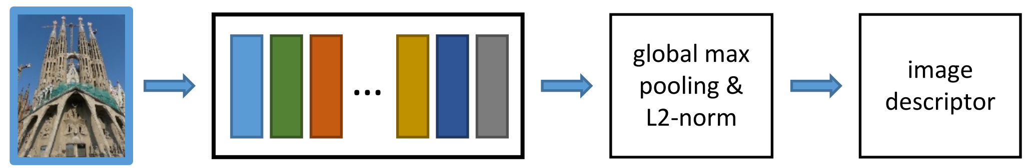
re-train with data relevant to your task

CNN Image Retrieval Learns from BoW: Unsupervised Fine-Tuning with Hard Examples

CNN Image Retrieval

compact image descriptors

Nearest Neighbor search



CNN Learning (Fine-Tuning)

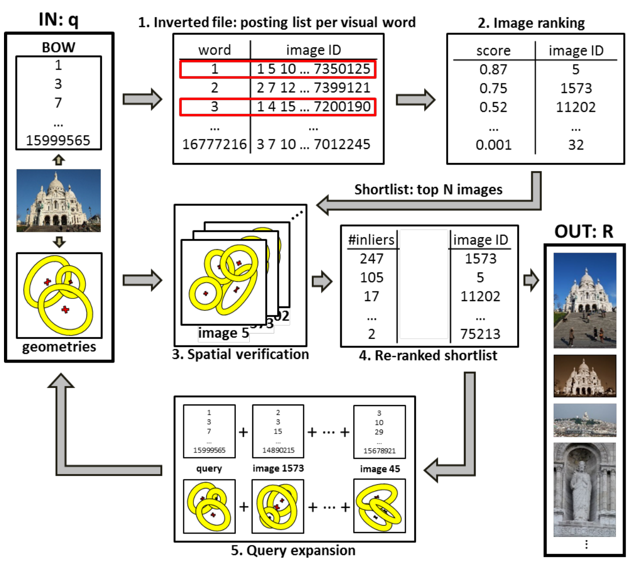
start with CNN trained for different but similar task (reasonable parameters)

re-train with data relevant to your task

Bag of Words

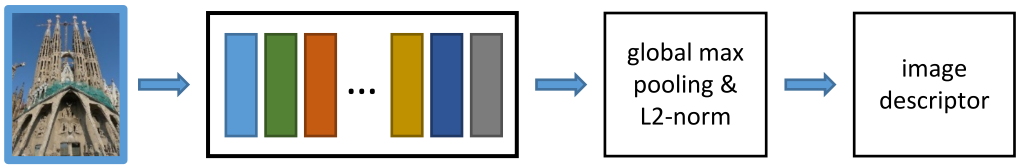
state-of-the-art retrieval performance

couples well with SfM



CNN Image Retrieval Learns from BoW: Unsupervised Fine-Tuning with Hard Examples

CNN Image Retrieval



- compact image descriptors
- Nearest Neighbor search

CNN Learning (Fine-Tuning)

- start with CNN trained for different but similar task (reasonable parameters)
- re-train with data relevant to your task

Bag of Words

- state-of-the-art retrieval performance
- couples well with SfM

Unsupervised training data generation

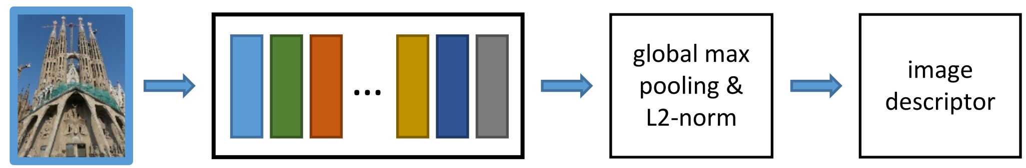
- no human interaction

CNN Image Retrieval Learns from BoW: Unsupervised Fine-Tuning with Hard Examples

CNN Image Retrieval

compact image descriptors

Nearest Neighbor search



CNN Learning (Fine-Tuning)

start with CNN trained for different but similar task (reasonable parameters)

re-train with data relevant to your task

Bag of Words

state-of-the-art retrieval performance

couples well with SfM

Unsupervised training data generation

no human interaction

Hard Examples

Instance Retrieval Challenges

- Significant viewpoint and/or scale change
- Significant illumination change
- Severe occlusions
- Visually similar but different objects

BoW: affine co-variant local features, invariant descriptors

CNN: lots of training examples

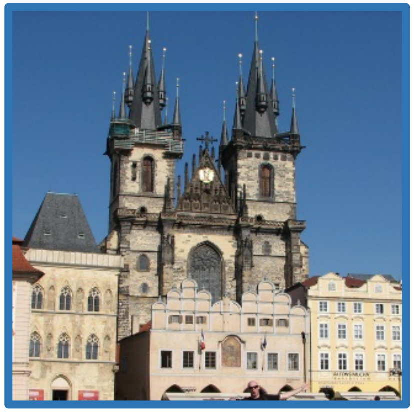
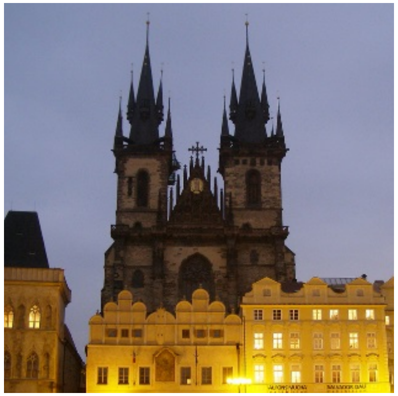


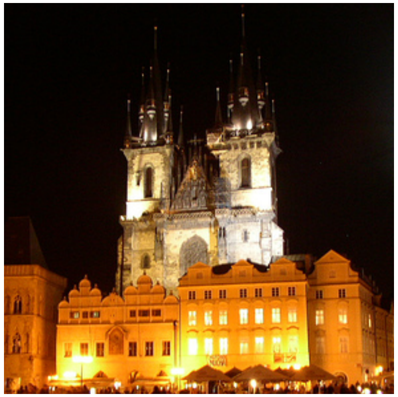


Instance Retrieval Challenges

- Significant viewpoint and/or scale change
- Significant illumination change
- Severe occlusions
- Visually similar but different objects

BoW: color-normalized feature descriptors
CNN: lots of training examples



Instance Retrieval Challenges

- Significant viewpoint and/or scale change
- Significant illumination change
- Severe occlusions
- Visually similar but different objects

BoW: locality of the features, geometric verification
CNN: lots of training examples

Instance Retrieval Challenges

Significant viewpoint and/or scale change

Significant illumination change

Severe occlusions

Visually similar but different objects

BoW: discriminability of the features, geometric verification

CNN: lots of training examples

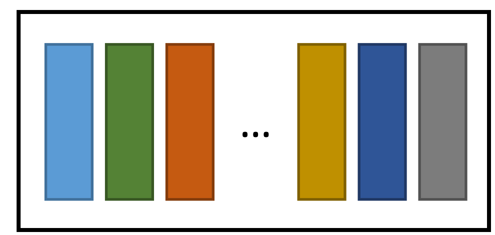



“Lots of Training Examples”

Large Internet
photo collection

Training

Image annotations



Convolutional Neural
Network (CNN)

“Lots of Training Examples”

Manual cleaning of the training data done by Researchers

Very expensive $$$$

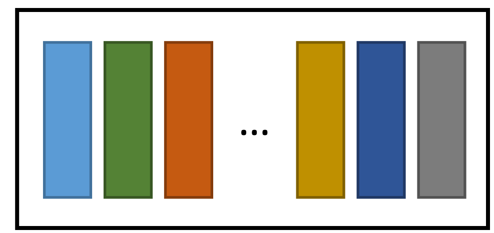
Large Internet photo collection

amazon mechanicalturk™

Artificial Artificial Intelligence

Not accurate

Expensive $$



Convolutional Neural Network (CNN)

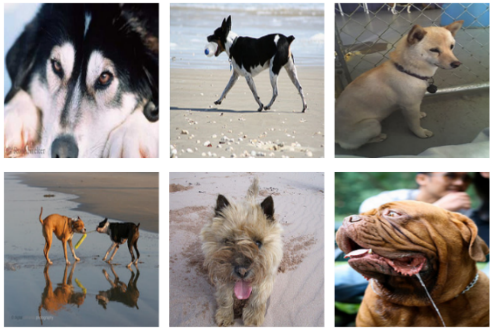
Automated extraction of training data

Very accurate

Free ~~$~~

Off-the-shelf CNN

- Target application: classification
- Training dataset: ImageNet
- Architecture: AlexNet & VGG



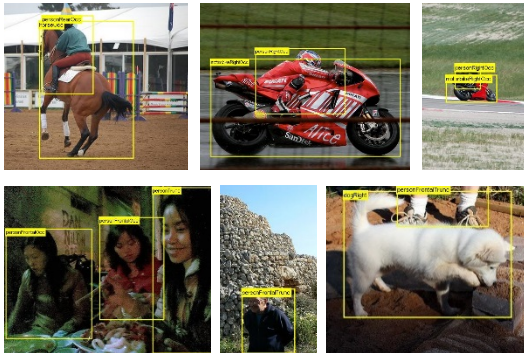
Images from ImageNet.org

- Directly applicable to other tasks

Fine-grain classification

Images from ImageNet.org

Object detection



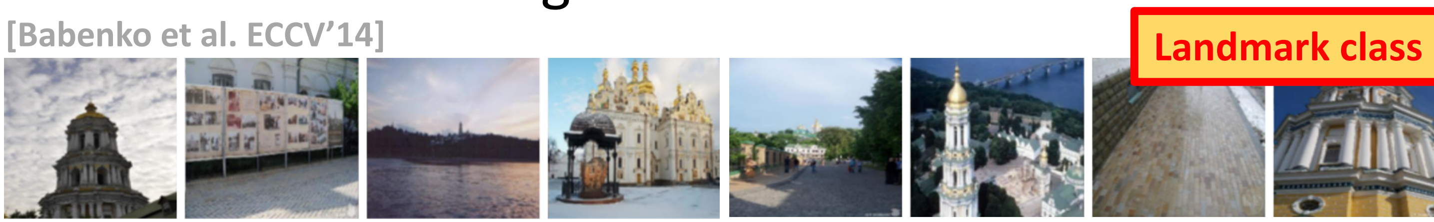
Images from PASCAL VOC 2012

Image retrieval

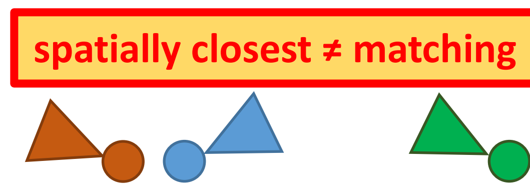
Annotations for CNN Image Retrieval

- CNN pre-trained for classification task used for retrieval
  [Gong et al. ECCV'14, Babenko et al. ICCV'15, Kalantidis et al. arXiv'15, Tolias et al. ICLR'16]

- Fine-tuned CNN using a dataset with landmark classes
  [Babenko et al. ECCV'14]



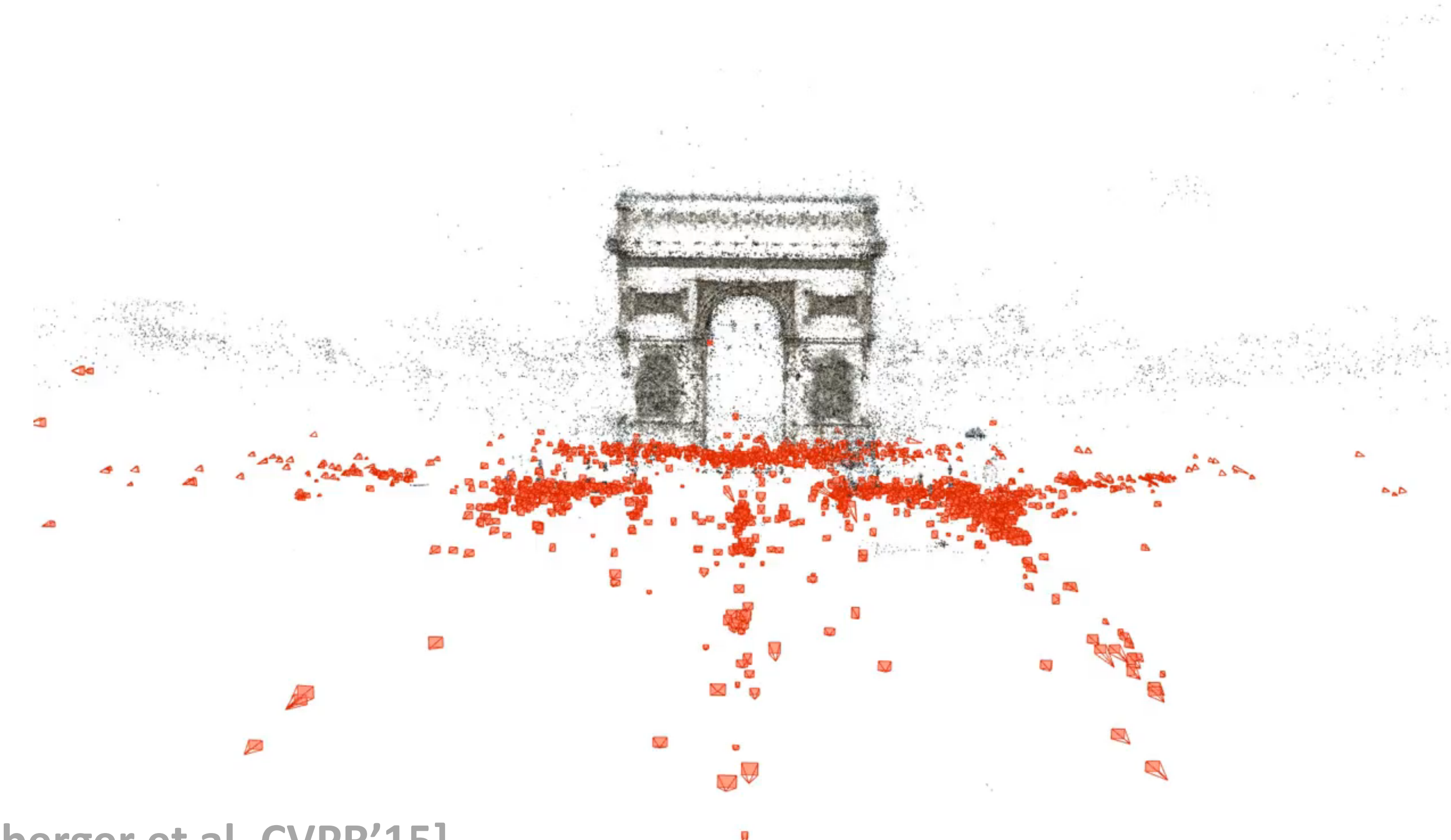
- NetVLAD: Weakly supervised fine-tuned CNN using GPS tags
  [Arandjelovic et al. CVPR'16]



- We propose: automatic annotations for CNN training

CNN learns from BoW – Training Data

Camera Orientation Known
Number of Inliers Known



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

7.4M images → 713 training 3D models

Hard Negative Examples

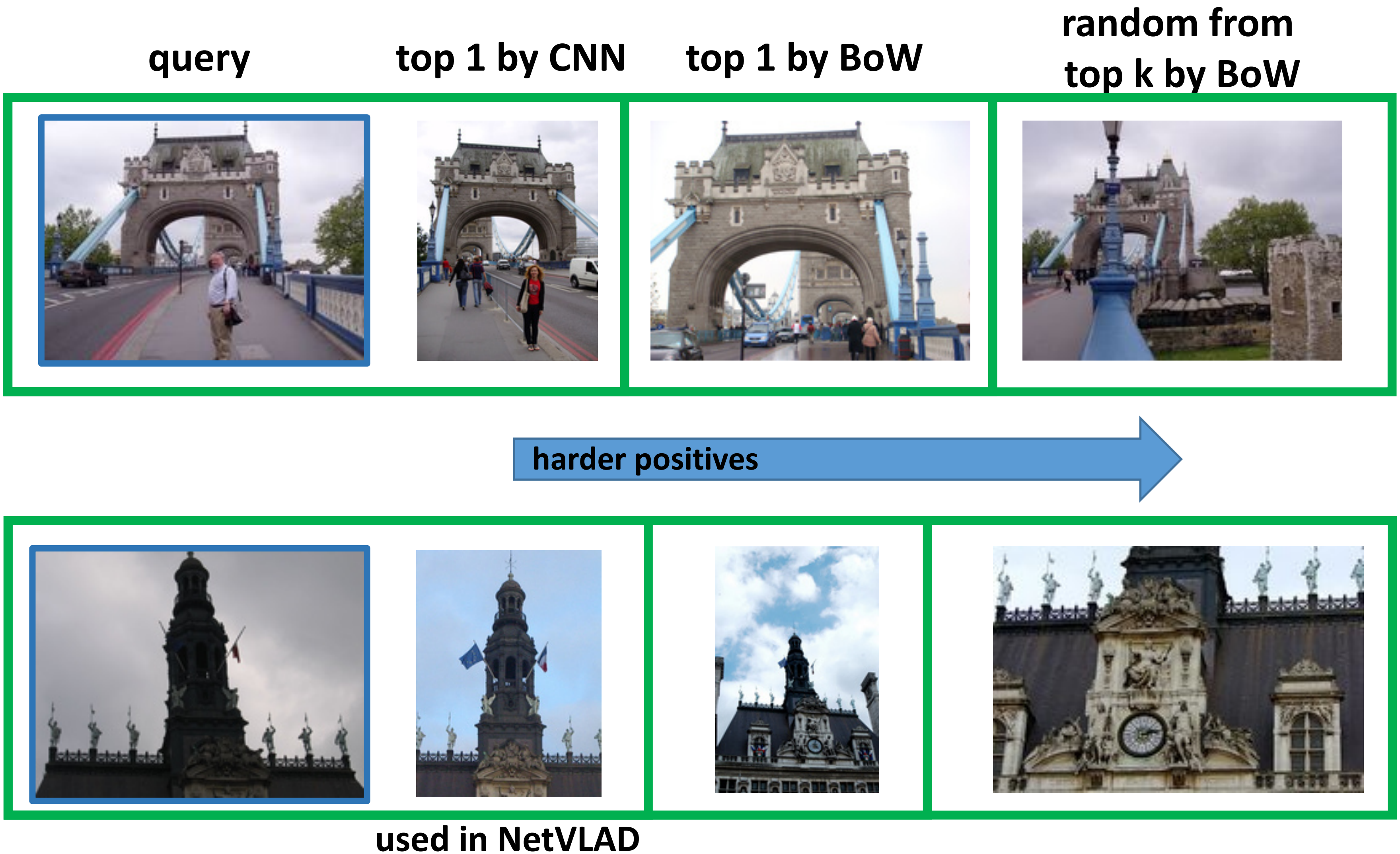
Negative examples: images from different 3D models than the query
Hard negatives: closest negative examples to the query
Only hard negatives: as good as using all negatives, but faster

increasing CNN descriptor distance to the query

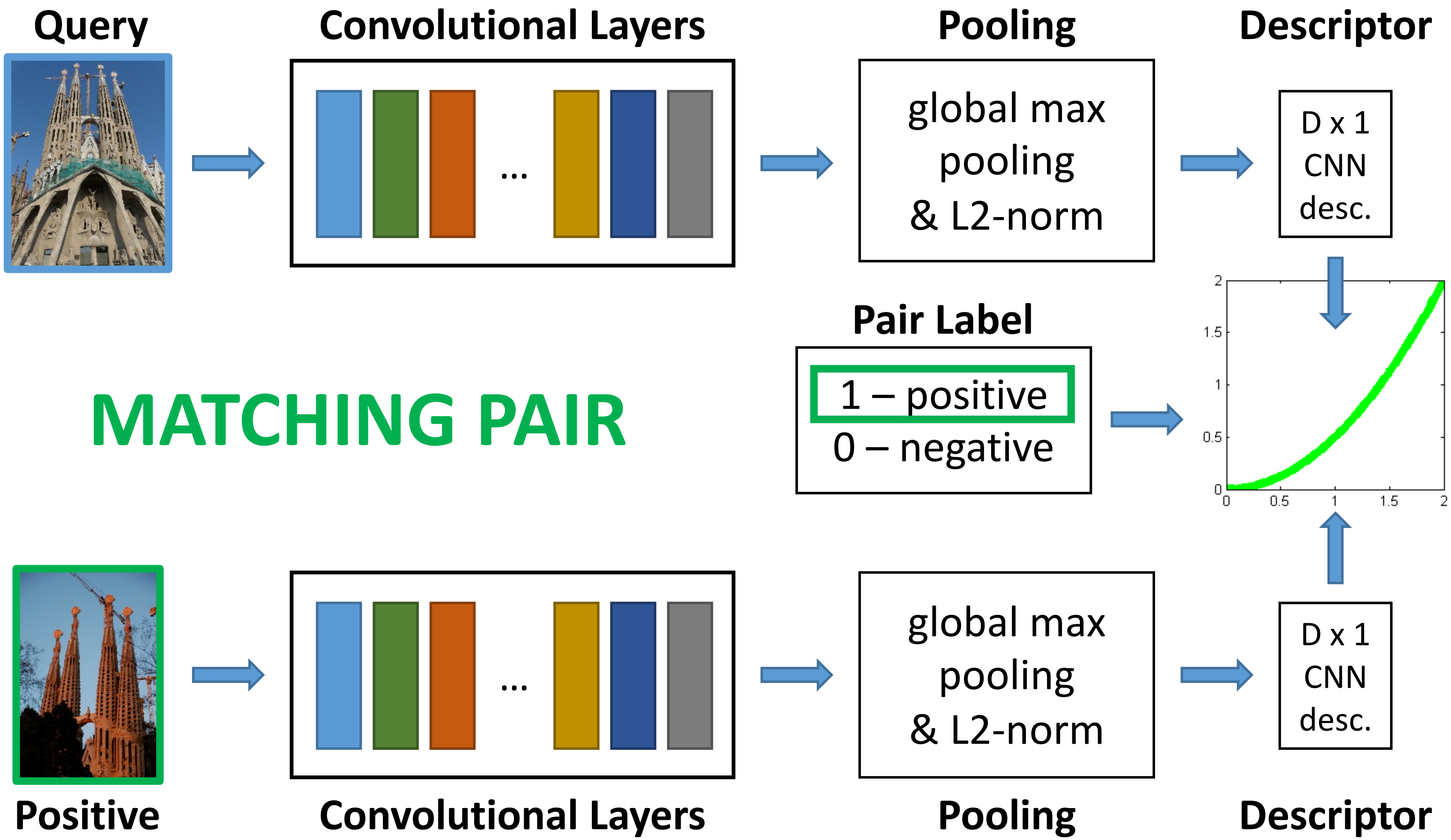
query	the most similar CNN descriptor	naive hard negatives top k by CNN	diverse hard negatives top k: one per 3D model

Hard Positive Examples

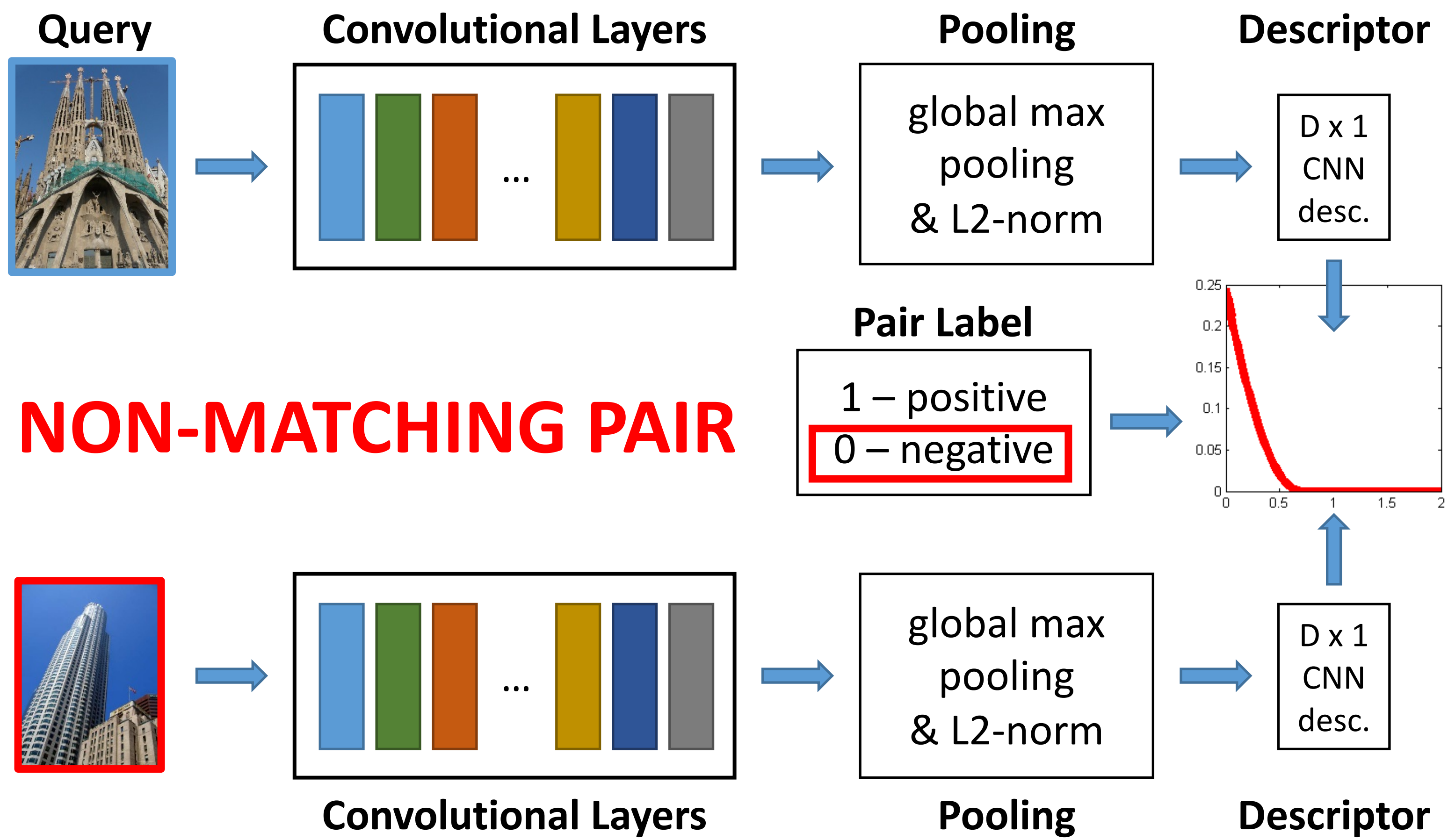
Positive examples: images that share 3D points with the query
Hard positives: positive examples not close enough to the query



CNN Siamese Learning



CNN Siamese Learning

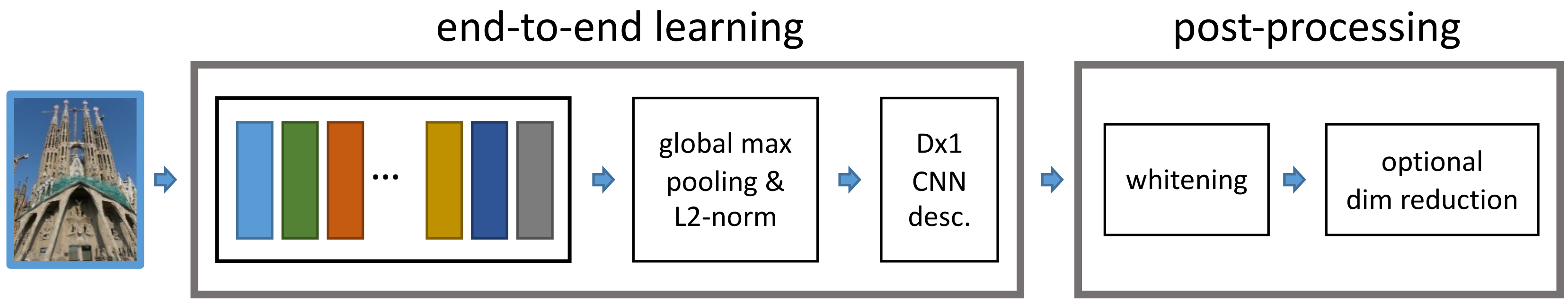


<u>Contrastive vs. Triplet loss</u>: Contrastive better with our data

Contrastive loss more strict, requires accurate training data

Triplet loss less sensitive to inaccurate annotation

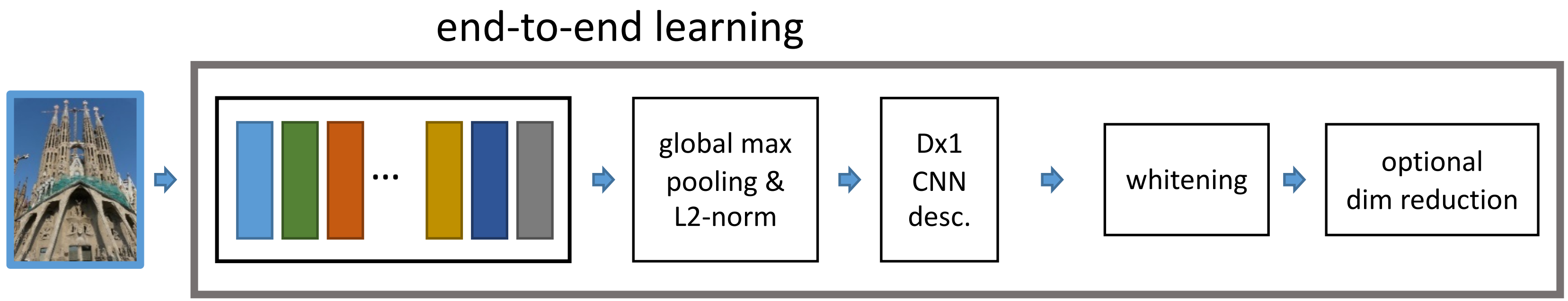
Whitening and dimensionality reduction



1. PCA_w – PCA of an independent set of descriptors
   [Babenko et al. ICCV'15, Tolias et al. ICLR'16]

2. L_w – We propose to learn whitening using labeled training data and linear discriminant projections
   [Mikolajczyk & Matas ICCV'07]

Whitening and dimensionality reduction

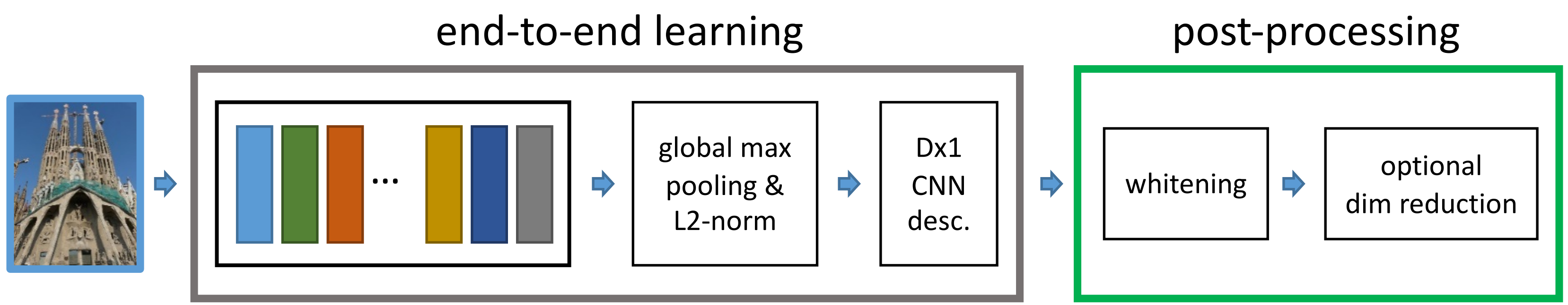


1. PCA_w – PCA of an independent set of descriptors
[Babenko et al. ICCV'15, Tolias et al. ICLR'16]

2. L_w – We propose to learn whitening using labeled training data and linear discriminant projections
[Mikolajczyk & Matas ICCV'07]

3. End-to-end Learning – Performs comparable or worse than L_w, while slowing down the convergence

Whitening and dimensionality reduction



1. PCA_w – PCA of an independent set of descriptors
[Babenko et al. ICCV'15, Tolias et al. ICLR'16]

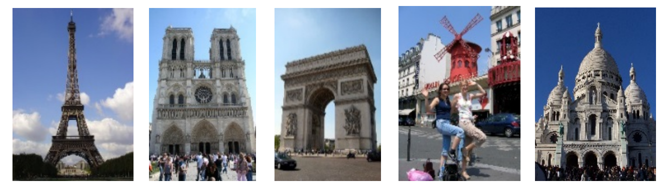
2. L_w – We propose to learn whitening using labeled training data and linear discriminant projections
[Mikolajczyk & Matas ICCV'07]

3. End-to-end Learning – Performs comparable or worse than L_w, while slowing down the convergence

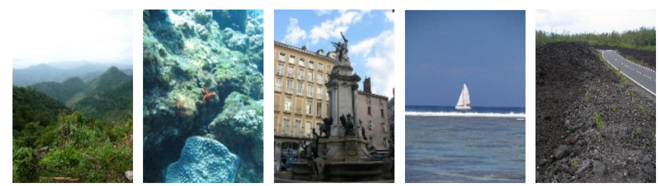
Experiments – datasets

- **Oxford 5k dataset**
  [Philbin et al. CVPR'07]

- **Paris 6k dataset**
  [Philbin et al. CVPR'08]



- **Holidays dataset**
  [Jegou et al. ECCV'10]



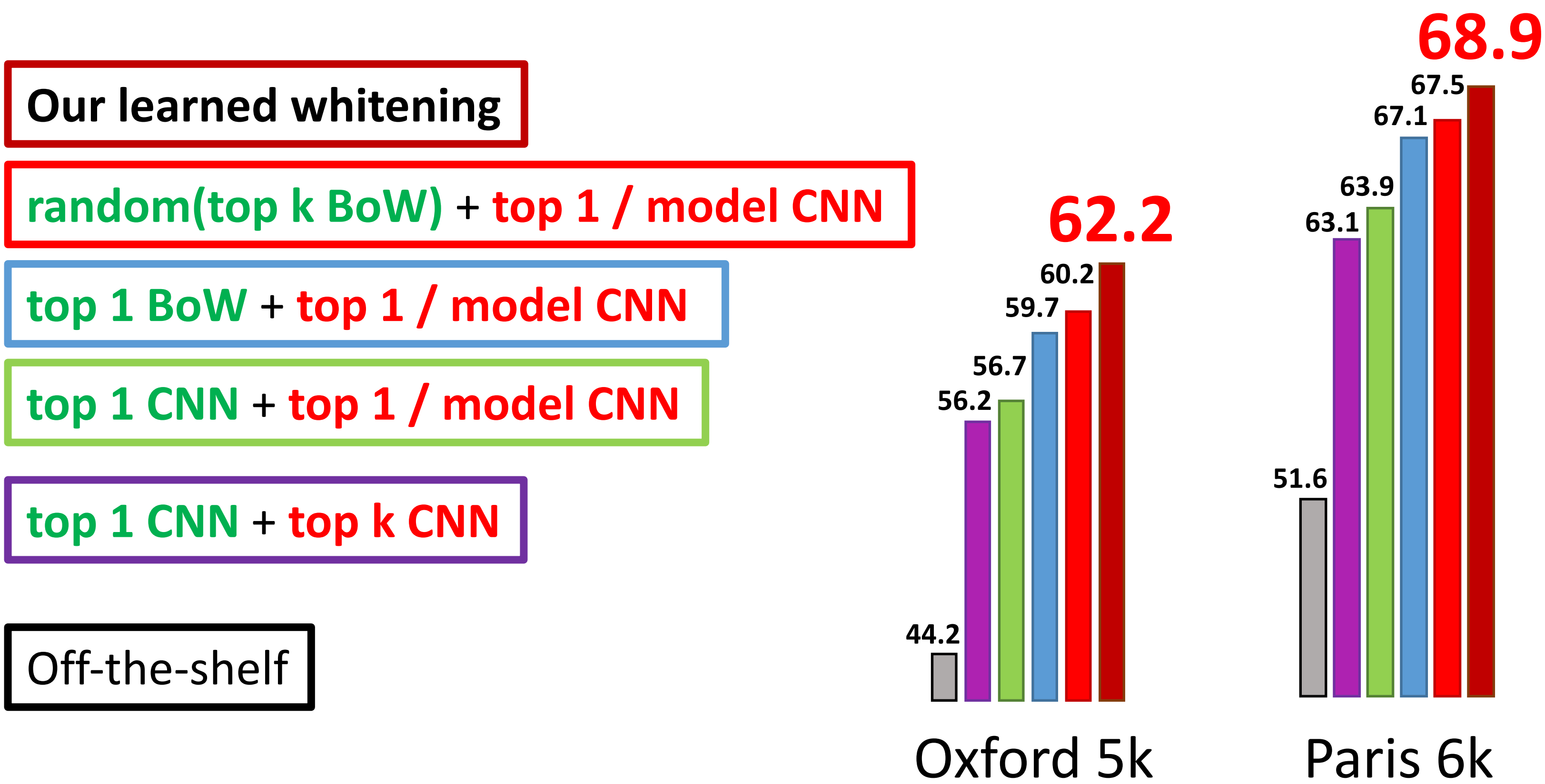
- **100k distractor dataset**
  [Philbin et al. CVPR'07]

Training 3D models do not contain any landmark from these datasets

- **Protocol:** mean Average Precision (mAP)

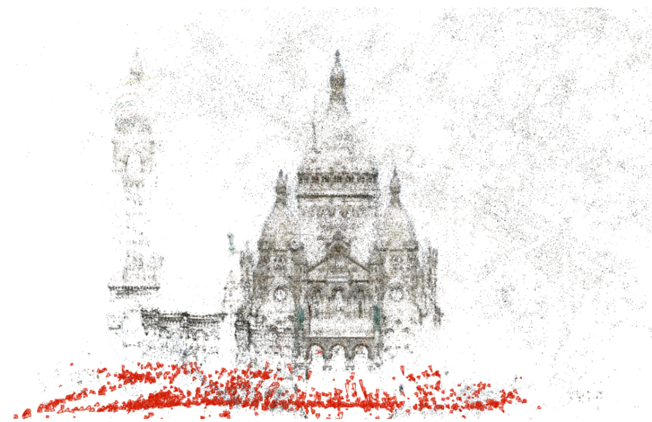
Experiments – Learning (AlexNet)

- Careful choice of **positive** and **negative** training images makes a difference



Experiments – Over-fitting and Generalization

- We added Oxford and Paris landmarks as 3D models and repeated fine-tuning



Only +0.3 mAP on average over all testing datasets

State-of-the-art

NetVLAD 256D

vs.

Our CNN 32D

Concurrent work:
[Gordo et al. ECCV'16]

63.5

69.2

Method		D	Oxf5k Crop$_\mathcal{I}$	Oxf5k Crop$_\mathcal{X}$	Oxf105k Crop$_\mathcal{I}$	Oxf105k Crop$_\mathcal{X}$	Par6k Crop$_\mathcal{I}$	Par6k Crop$_\mathcal{X}$	Par106k Crop$_\mathcal{I}$	Par106k Crop$_\mathcal{X}$	Hol	Hol 101k
Compact representations												
mVoc/BoW [11]		128	48.8	–	41.4	–	–	–	–	–	65.6	–
Neural codes† [14]	(fA)	128	–	**55.7**	–	**52.3**	–	–	–	–	**78.9**	–
MAC‡	(V)	128	53.5	**55.7**	43.8	45.6	69.5	**70.6**	53.4	**55.4**	72.6	**56.7**
CroW [24]	(V)	128	**59.2**	–	**51.6**	–	**74.6**	–	**63.2**	–	–	–
★ MAC	(fV)	128	**75.8**	**76.8**	**68.6**	**70.8**	77.6	78.8	68.0	69.0	73.2	58.8
★ R-MAC	(fV)	128	72.5	76.7	64.3	69.7	**78.5**	**80.3**	**69.3**	**71.2**	**79.3**	**65.2**
MAC‡	(V)	256	54.7	56.9	45.6	47.8	71.5	72.4	55.7	**57.3**	76.5	**61.3**
SPoC [23]	(V)	256	–	53.1	–	**50.1**	–	–	–	–	80.2	–
R-MAC [25]	(A)	256	56.1	–	47.0	–	72.9	–	60.1	–	–	–
CroW [24]	(V)	256	**65.4**	–	**59.3**	–	**77.9**	–	**67.8**	–	83.1	–
NetVlad [35]	(V)	256	[illegible]	[illegible]	[illegible]	–	–	67.7	–	–	**86.0**	–
NetVlad [35]	(fV)	256	[illegible]	[illegible]	[illegible]	–	–	**73.5**	–	–	84.3	–
★ MAC	(fA)	256	[illegible]	[illegible]	[illegible]	58.0	68.9	72.2	54.7	58.5	76.2	63.8
★ R-MAC	(fA)	256	62.5	68.9	53.2	61.2	74.4	76.6	61.8	64.8	81.5	**70.8**
★ MAC	(fV)	256	**77.4**	**78.2**	**70.7**	**72.6**	80.8	81.9	72.2	73.4	77.3	62.9
★ R-MAC	(fV)	256	74.9	**78.2**	67.5	72.1	**82.3**	**83.5**	**74.1**	**75.6**	81.4	69.4
MAC‡	(V)	512	56.4	**58.3**	47.8	**49.2**	72.3	**72.6**	58.0	**59.1**	76.7	**62.7**
R-MAC [25]	(V)	512	66.9	–	61.6	–	**83.0**	–	**75.7**	–	–	–
CroW [24]	(V)	512	**68.2**	–	**63.2**	–	79.6	–	71.0	–	84.9	–
★ MAC	(fV)	512	**79.7**	80.0	**73.9**	**75.1**	82.4	82.9	74.6	75.3	79.5	67.0
★ R-MAC	(fV)	512	77.0	**80.1**	69.2	74.1	**83.8**	**85.0**	**76.4**	**77.9**	82.5	**71.5**
Extreme short codes												
Neural codes† [14]	(fA)	16	–	**41.8**	–	**35.4**	–	–	–	–	**60.9**	–
★ MAC	(fV)	16	**56.2**	**57.4**	**45.5**	**47.6**	57.3	62.9	43.4	48.5	51.3	25.6
★ R-MAC	(fV)	16	46.9	52.1	37.9	41.6	**58.8**	**63.2**	**45.6**	**49.6**	54.4	**31.7**
Neural codes† [14]	(fA)	32	[illegible]	[illegible]	[illegible]	**46.7**	–	–	–	–	**72.9**	–
★ MAC	(fV)	32	[illegible]	[illegible]	[illegible]	**59.5**	**63.9**	**69.5**	51.6	**56.3**	62.4	41.8
★ R-MAC	(fV)	32	[illegible]	[illegible]	[illegible]	55.1	**63.9**	67.4	**52.7**	55.8	68.0	**49.6**
Re-ranking (R) and query expansion (QE)												
BoW(1M)+QE [6]		–	82.7	–	76.7	–	80.5	–	71.0	–	–	–
BoW(16M)+QE [50]		–	84.9	–	79.5	–	82.4	–	77.3	–	–	–
HQE(65k) [8]		–	**88.0**	–	**84.0**	–	82.8	–	–	–	–	–
R-MAC+R+QE [25]	(V)	512	77.3	–	73.2	–	**86.5**	–	**79.8**	–	–	–
CroW+QE [24]	(V)	512	72.2	–	67.8	–	85.5	–	79.7	–	–	–
★ MAC+R+QE	(fV)	512	85.0	**85.4**	81.8	**82.3**	**86.5**	**87.0**	78.8	79.6	–	–
★ R-MAC+R+QE	(fV)	512	82.9	84.5	77.9	80.4	85.6	86.4	78.3	**79.7**	–	–

Teacher vs. Student

Method	Oxf5k	Oxf105k	Par6k	Par106k
BoW(16M)+R+QE	**84.9**	**79.5**	**82.4**	**77.3**
CNN(512D)	79.7	73.9	**82.4**	74.6
CNN(512D)+R+QE	**85.0**	**81.8**	**86.5**	**78.8**

Our CNN with re-ranking (R) and query expansion(QE) surpasses its teacher on all datasets!!!

Teacher vs. Student

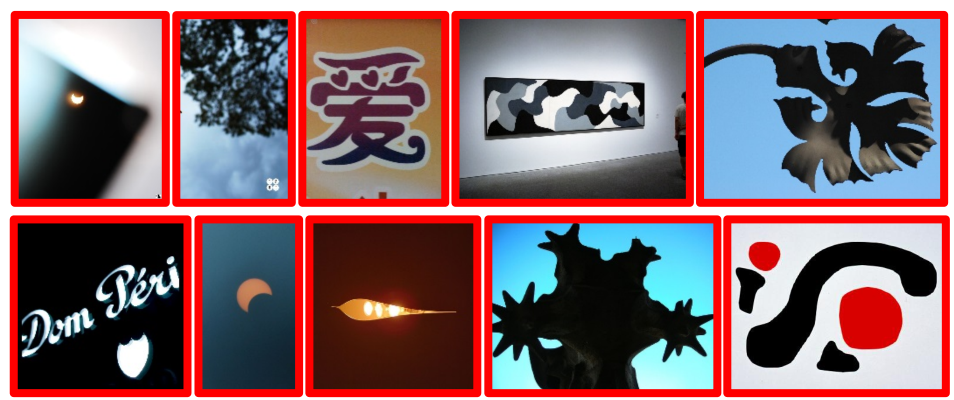
top 10 (correct | incorrect)

query

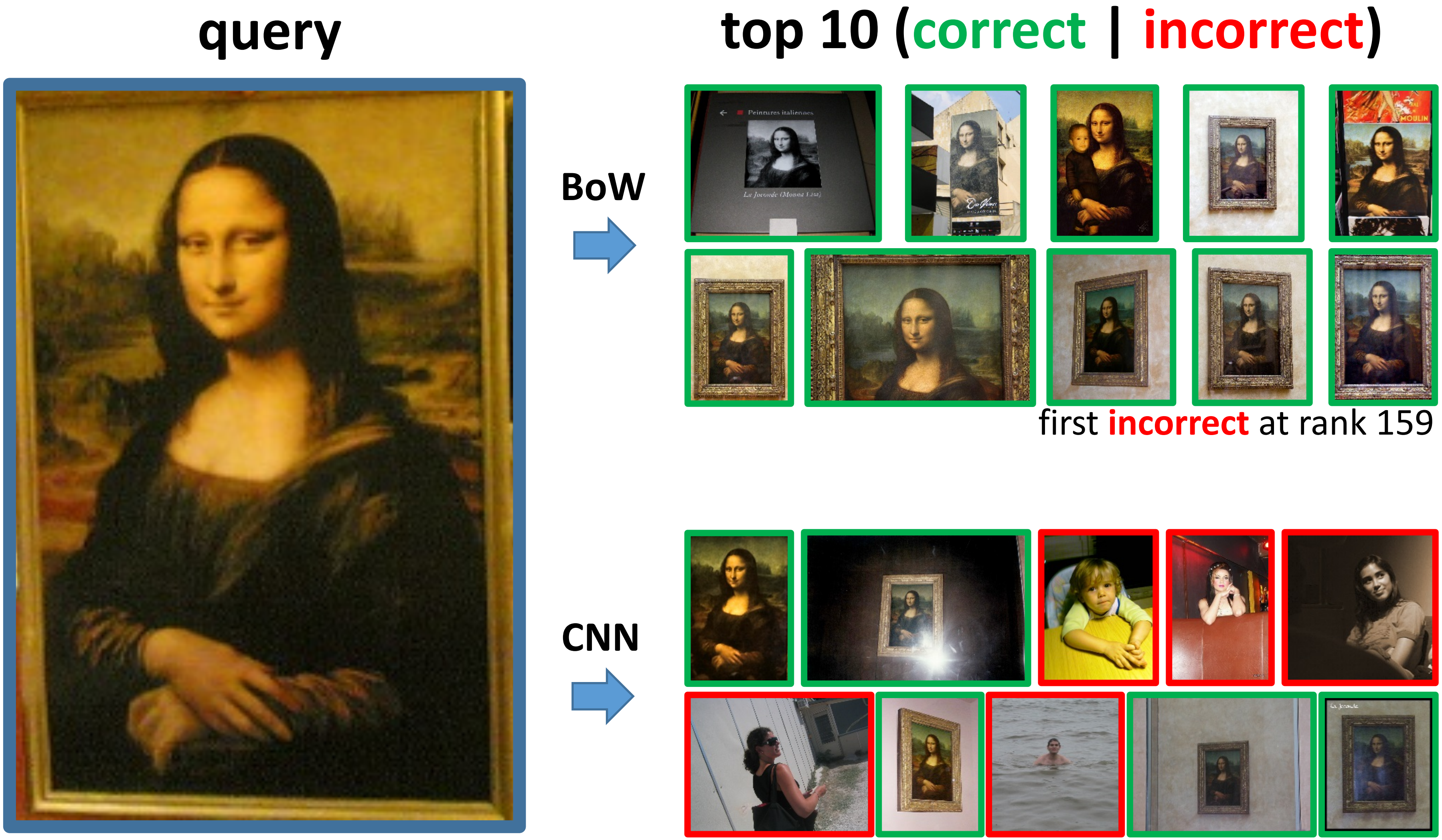
BoW

first incorrect at rank 127

CNN



Teacher vs. Student



Teacher vs. Student

Fine-tuning might not be enough

Conclusions

- We propose a method to generate the necessary “lots of training examples” without any human interaction
- Strong supervision for hard negative, hard positive mining, and supervised whitening
- Data and trained networks available at: cmp.felk.cvut.cz/~radenfil/projects/siamac.html
- For more details about the paper visit **Poster O-1A-01**