We Don’t Need No Annotation

(Efficient Training for Image Retrieval)

Ondra Chum

Visual Recognition Group

Department of Cybernetics, Faculty of Electrical Engineering

CTU in Prague

Outline

Algorithmic supervision for CNN training
(local features based methods)

- CNN fine-tuning for efficient image retrieval
- Sketch based image retrieval with CNN descriptors

Unsupervised metric learning from data manifolds

CNN fine-tuning for image retrieval

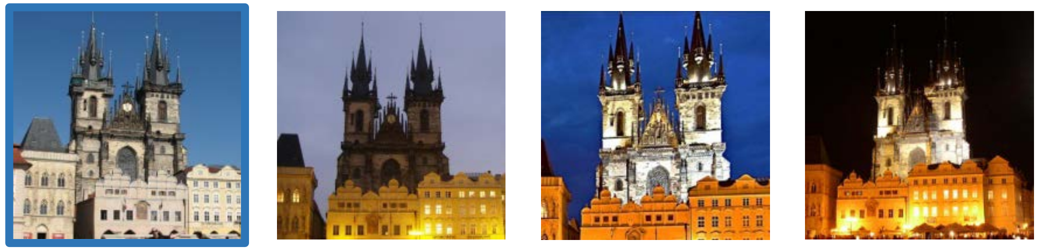
Filip Radenović

Giorgos Tolias

F. Radenovic, G. Tolias and O. Chum, CNN Image Retrieval Learns from BoW: Unsupervised Fine-Tuning with Hard Examples, In ECCV 2016

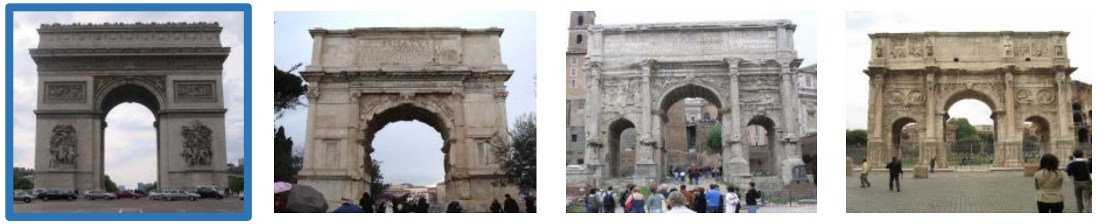
Image Retrieval Challenges

Significant viewpoint and/or scale change



Significant illumination change

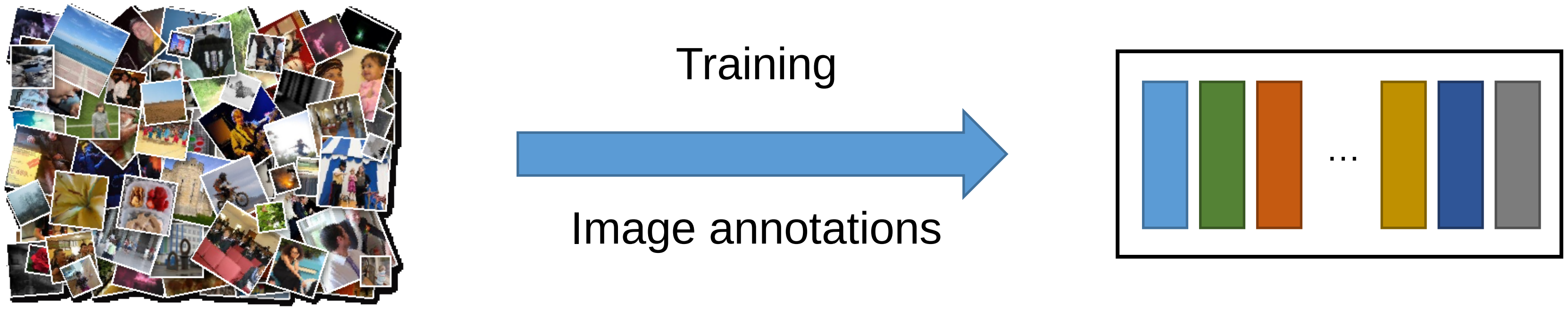
Severe occlusions



Visually similar but different objects

Old school:	local features, photometric normalization, geometric constraints
CNNs:	lots of training data, provides image embedding, nearest neighbor search

Lots of Training Examples



Large Internet photo collection

Convolutional Neural Network (CNN)

Lots of Training Examples

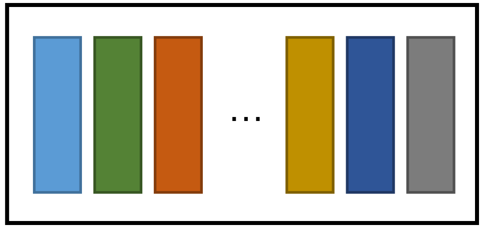
Manual cleaning of the training data done by Researchers

Very expensive $$$$

Large Internet photo collection

amazon mechanicalturk™
Artificial Artificial Intelligence

Not accurate
Not free $



Convolutional Neural Network (CNN)

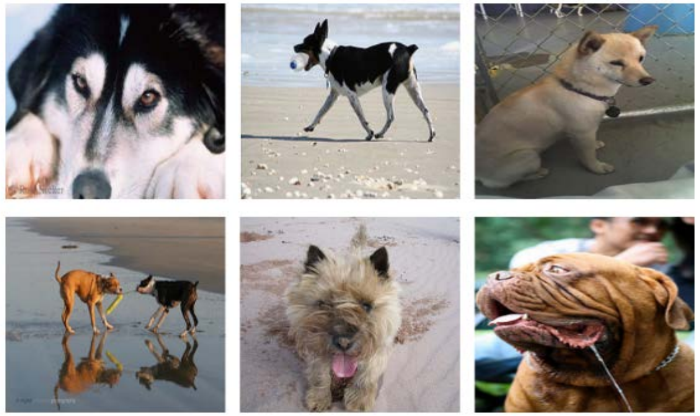
Automated extraction of training data

Accurate
Free ~~$~~

CNN Image Retrieval

- Image representation created from CNN activations of a network pre-trained for classification task

  [Gong et al. ECCV'14, Razavian et al. arXiv'14, Babenko et al. ICCV'15, Kalantidis et al. arXiv'15, Tolias et al. ICLR'16]



Images from ImageNet.org

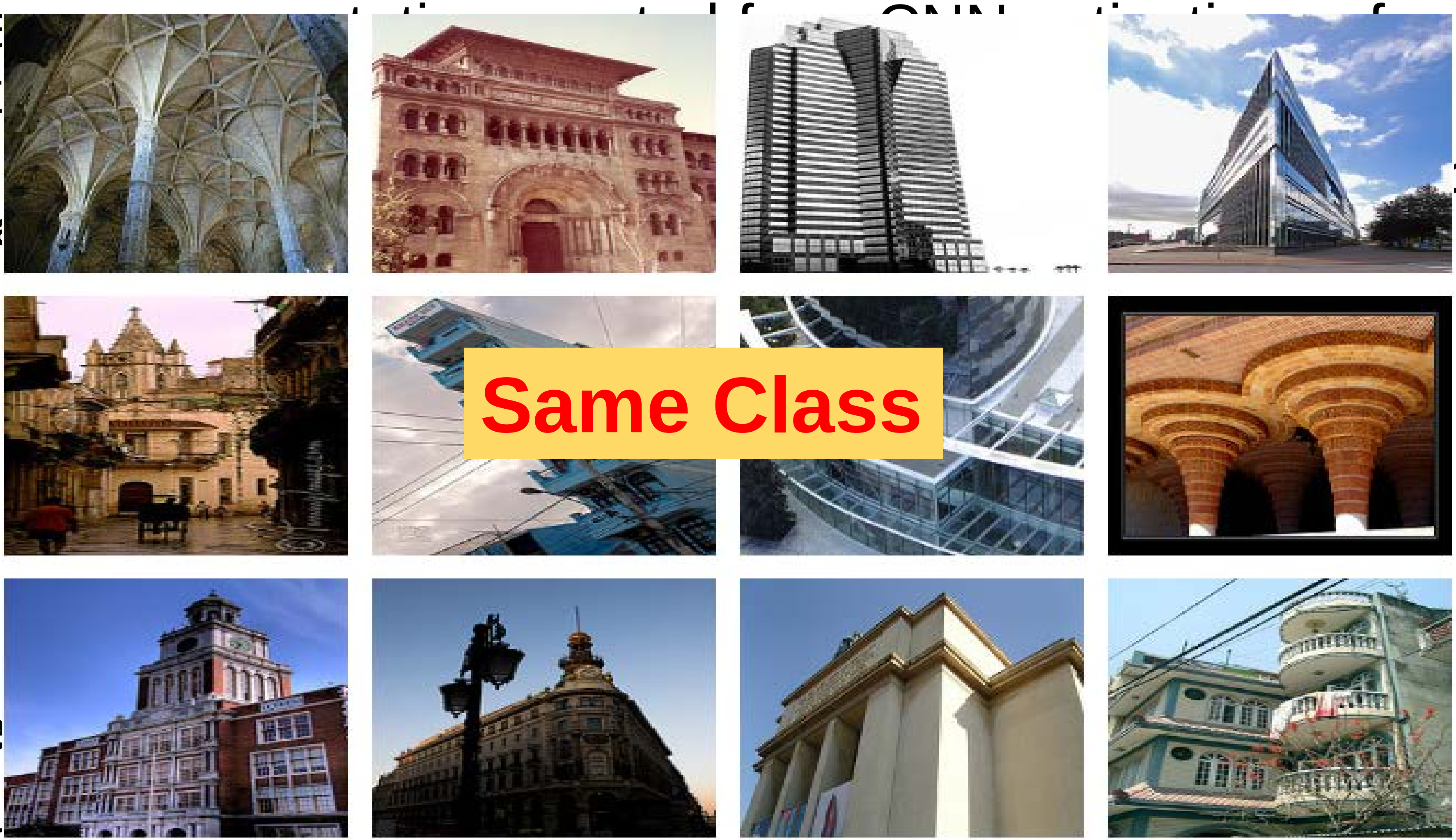
\+ Retrieval accuracy suggests generalization of CNNs

\- Trained for image classification, **NOT** retrieval task

CNN Image Retrieval

- Ima… net…

  [Go… 15, Kala…



\+ Re…

\- Trained for image classification, **NOT** retrieval task

CNN Image Retrieval

- CNN network re-trained using a dataset that contains landmarks and buildings as object classes.

  [Babenko et al. ECCV'14]

+ Training dataset closer to the target task

- Final metric different to the one actually optimized

- Constructing training datasets requires manual effort

CNN Image Retrieval

• CN… …ks and
bu…

[Ba…

+ Tr…

- Fi…

- Constructing training datasets requires manual effort

Image from [Babenko et al. ECCV'14]

CNN Image Retrieval

- NetVLAD: end-to-end fine-tuning for image retrieval. Geo-tagged dataset for weakly supervised fine-tuning.

  [Arandjelovic et al. CVPR’16]

+ Training dataset corresponds to the target task

+ Final metric corresponds to the one actually optimized

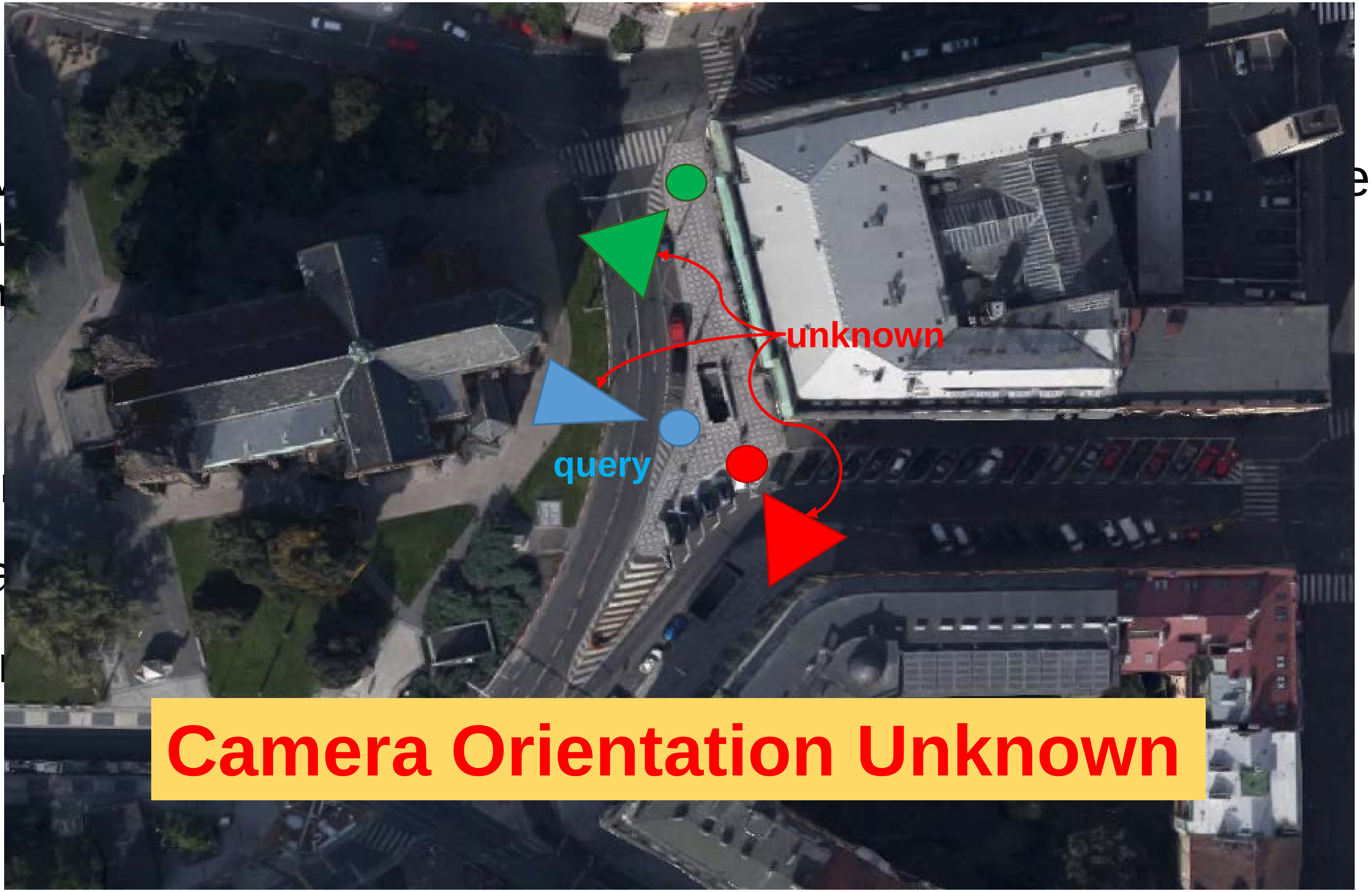
- Training dataset requires geo-tags

CNN Image Retrieval

- NetV… …ed data…

  [Aran…

+ Trai…
+ Fina…
- Trai…



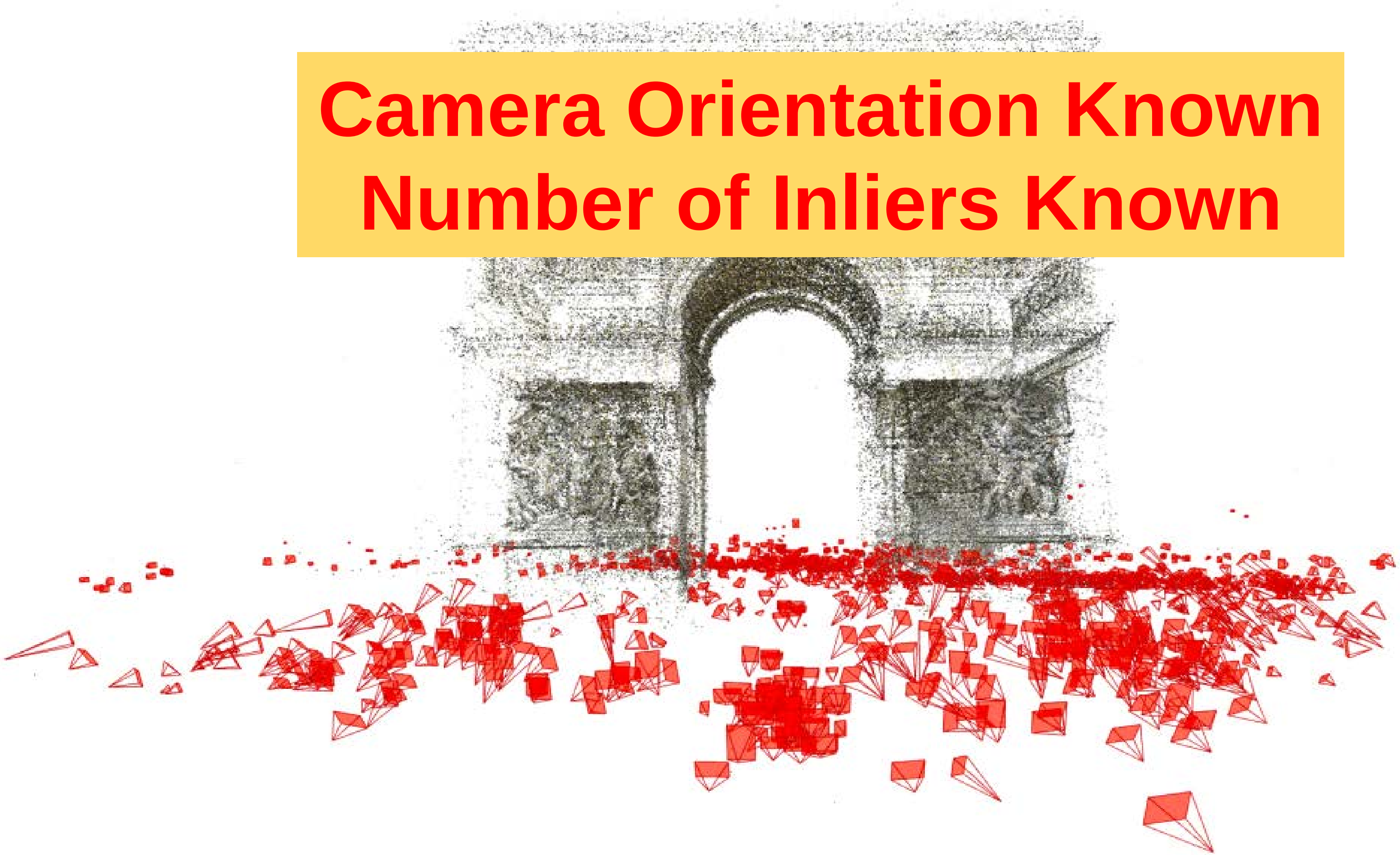
CNN learns from BoW – Training Data

Inp

1. s

2. e

Ou



Camera Orientation Known
Number of Inliers Known

Hard Negative Examples

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

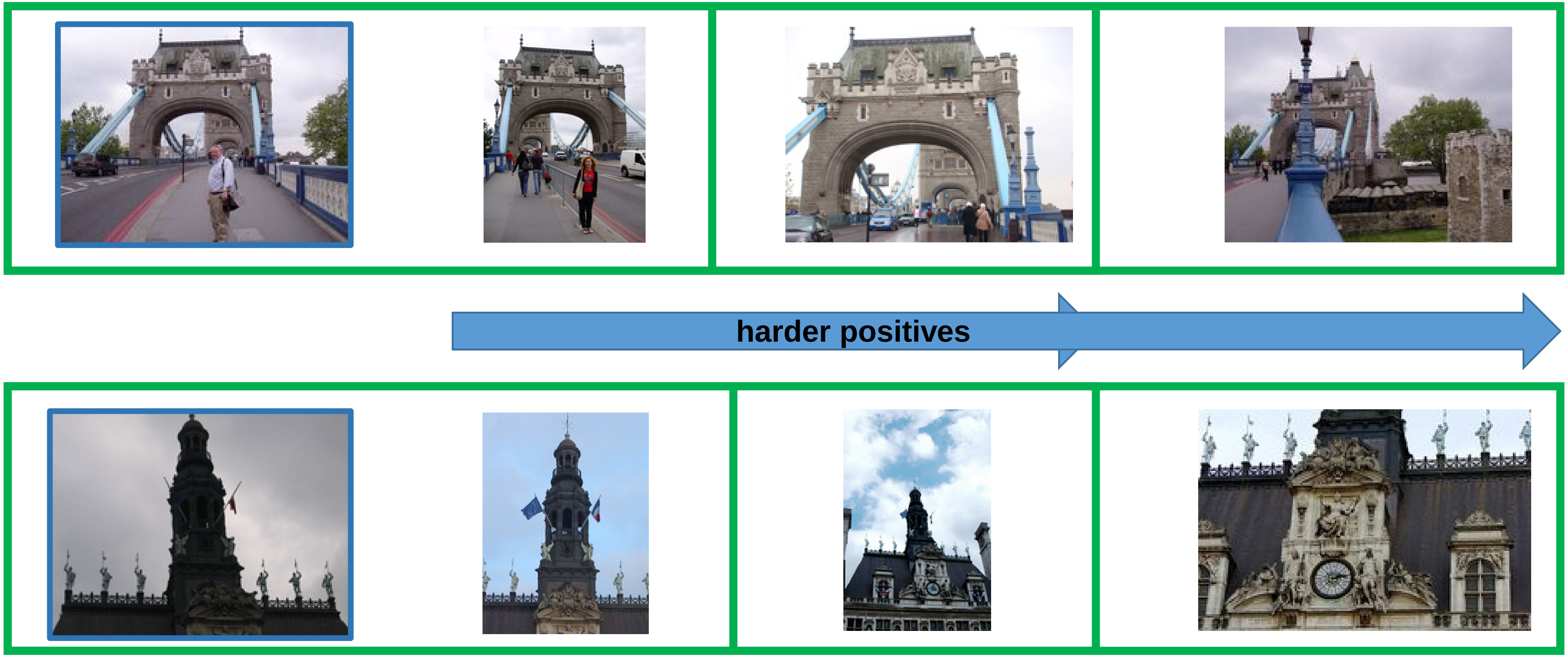
increasing CNN descriptor distance to the anchor

anchor	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 anchor
Hard positives: positive examples not close enough to the anchor

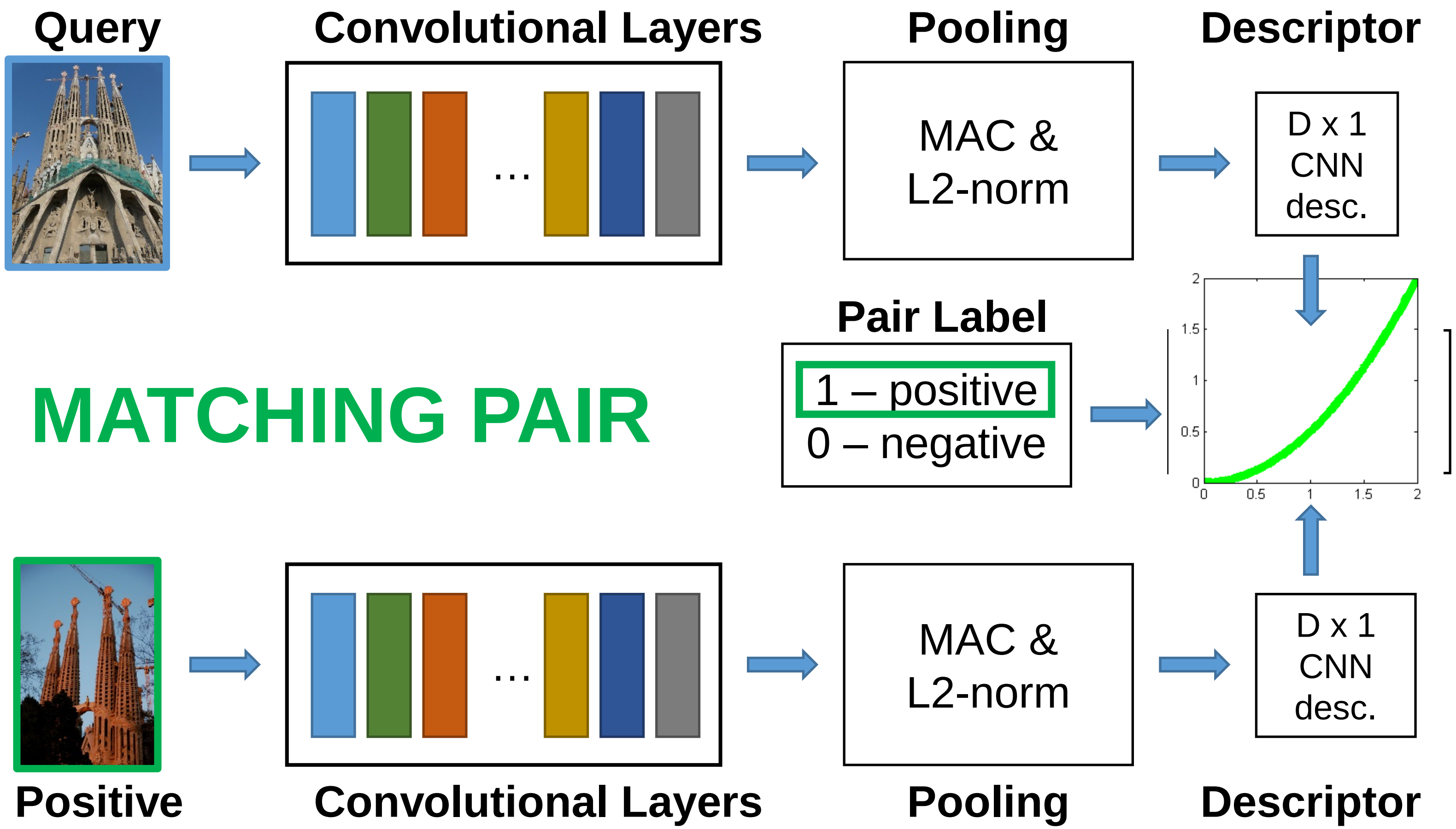
anchor | **top 1 by CNN** | **top 1 by BoW** | **random from top k by BoW**



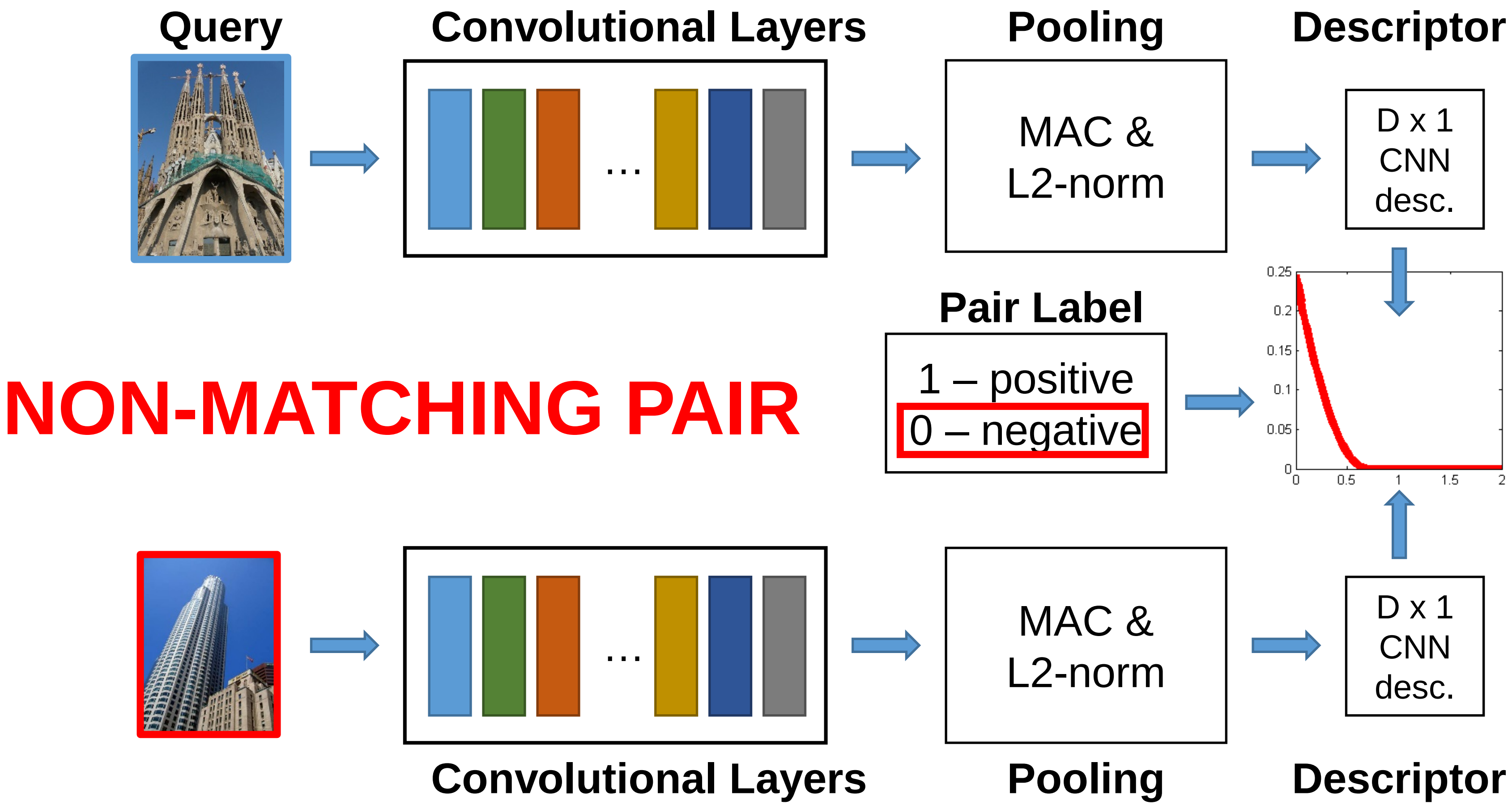
used in NetVLAD

F. Radenovic, G. Tolias and O. Chum, CNN Image Retrieval Learns from BoW: Unsupervised Fine-Tuning with Hard Examples, In ECCV 2016

CNN Siamese Learning

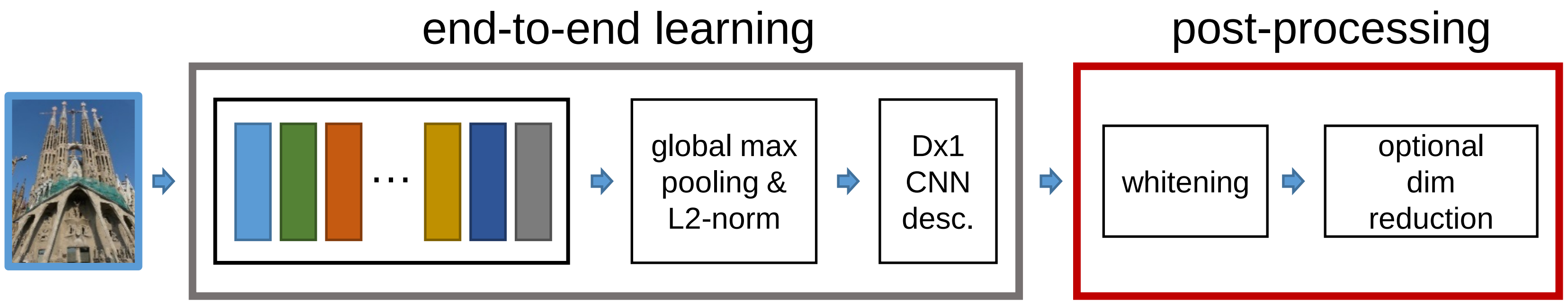


CNN Siamese Learning



F. Radenovic, G. Tolias and O. Chum, CNN Image Retrieval Learns from BoW: Unsupervised Fine-Tuning with Hard Examples, In ECCV 2016

Component Contributions (AlexNet)



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

MAC: learned whitening

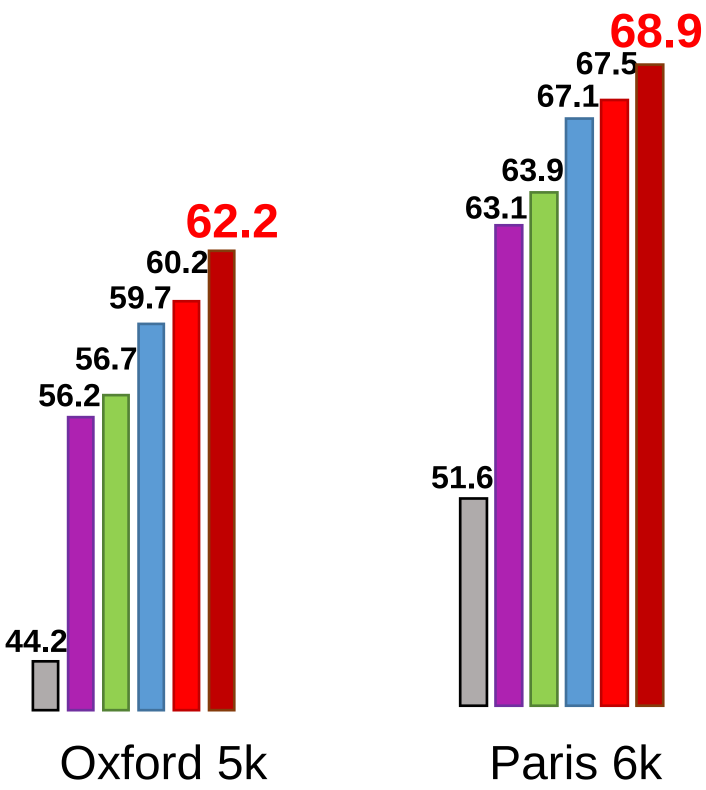
MAC: random(top k BoW) + top 1 / model CNN

MAC: top 1 BoW + top 1 / model CNN

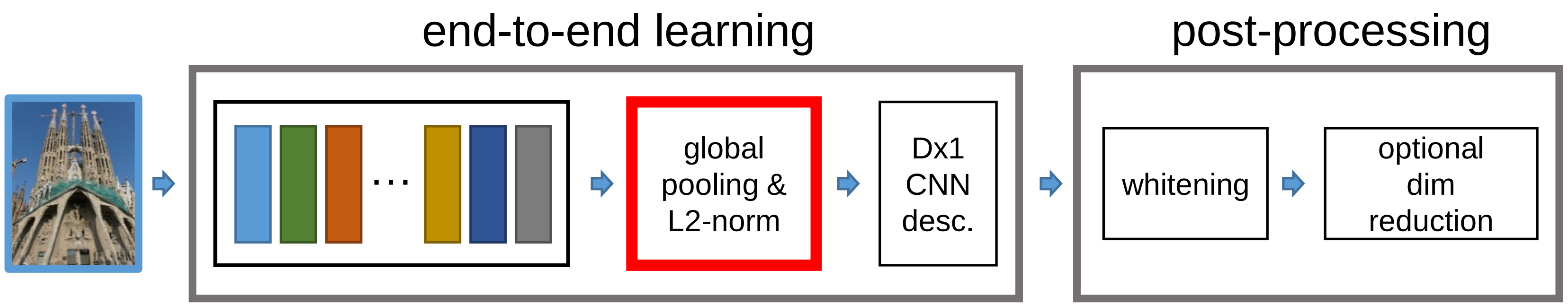
MAC: top 1 CNN + top 1 / model CNN

MAC: top 1 CNN + top k CNN

MAC: off-the-shelf



Global Pooling



MAC max pooling **M**aximum **A**ctivations of **C**onvolutions [Tolias et al. ICLR'16]

SPoC sum pooling **S**um-**Po**oled **C**onvolutional [Babenko et al. ICCV'15]

GeM generalized mean pooling **Ge**neralized **M**ean

p = 1
average pooling ← $$\left(\frac{1}{n}\sum_{i=1}^{n} x_i^p\right)^{\frac{1}{p}}$$ → p = inf
max pooling

[Radenovic, Tolias, Chum: TPAMI 2018]

Component Contributions (AlexNet)

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

GeM: learned whitening

GeM: random(top k BoW) + top 1 / model CNN

MAC: learned whitening

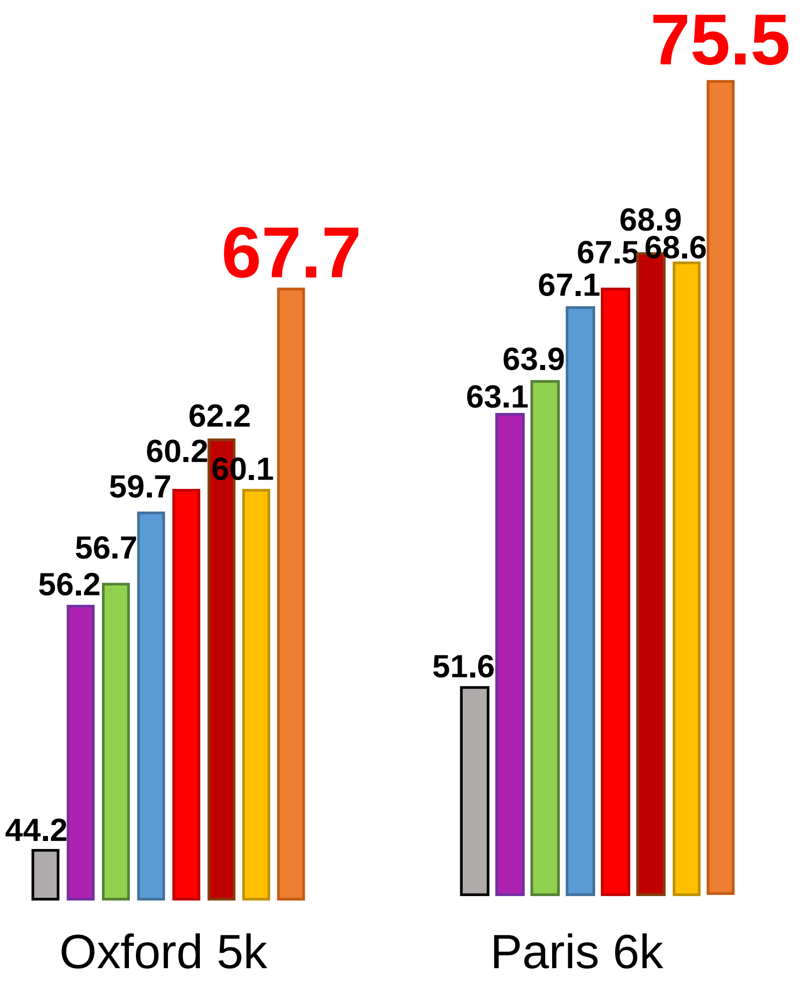
MAC: random(top k BoW) + top 1 / model CNN

MAC: top 1 BoW + top 1 / model CNN

MAC: top 1 CNN + top 1 / model CNN

MAC: top 1 CNN + top k CNN

MAC: off-the-shelf



Teacher vs. Student (VGG)

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

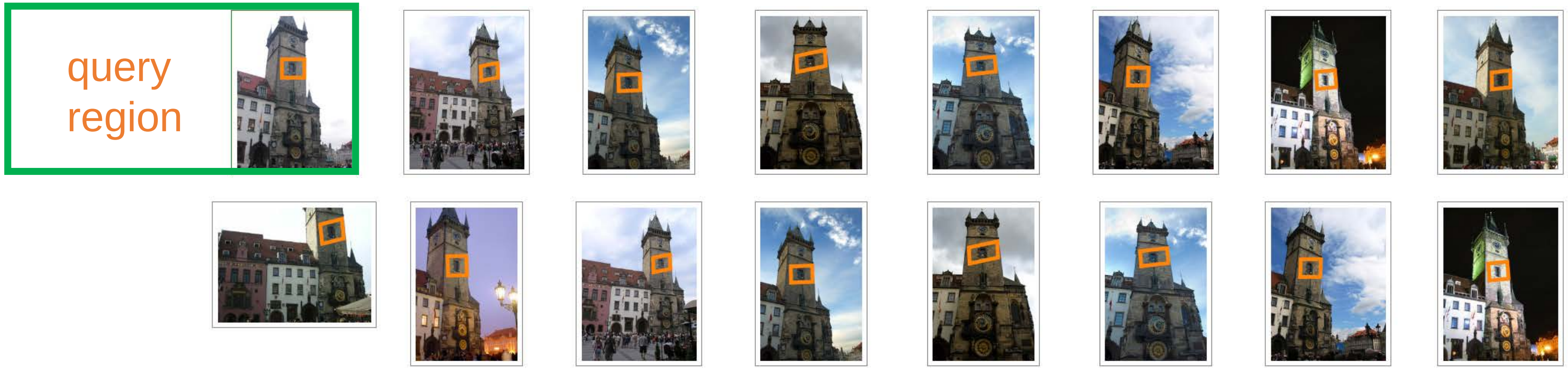
Teacher vs. Student (VGG)

Method	Oxf5k	Oxf105k	Par6k	Par106k
BoW(16M)+R+QE	**84.9**	**79.5**	**82.4**	**77.3**
CNN-MAC(512D)	79.7	73.9	**82.4**	74.6
CNN-GeM(512D)	**86.4**	**81.3**	**88.1**	**81.7**
CNN-GeM(512D)+QE	**90.7**	**88.6**	**92.2**	**88.0**

Our CNN with GeM layer surpasses
its teacher on all datasets!!! **BUT…**

Teacher vs. Student for small objects

CNN



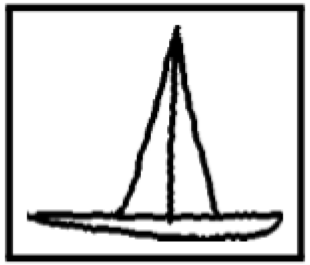
BoW+geometry

CNN fine-tuning for sketch-based image retrieval

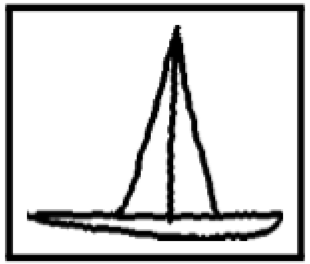
Filip Radenović

Giorgos Tolias

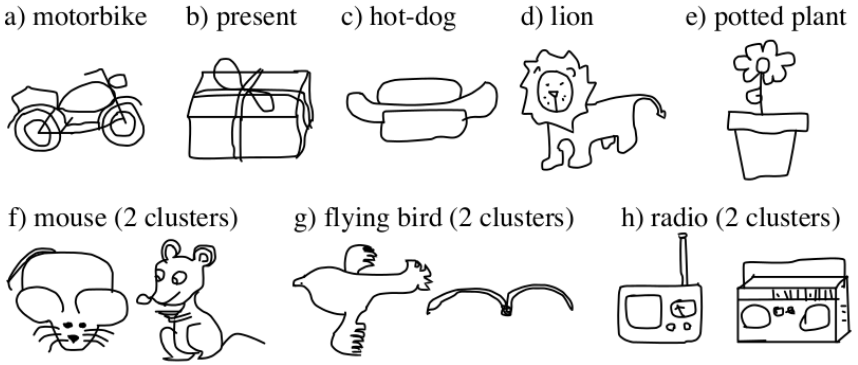
Sketch-based Image Retrieval

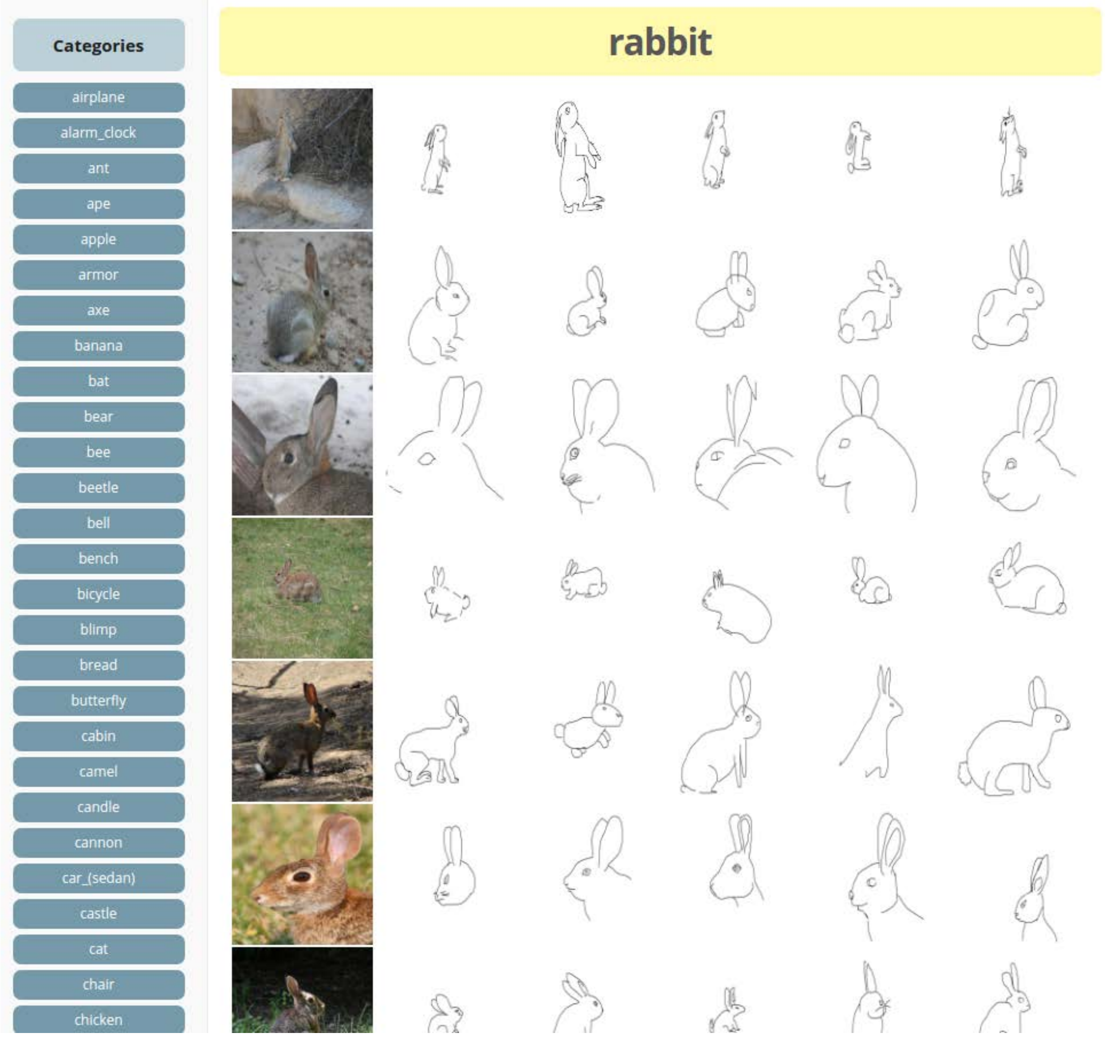


Sketch-based Image Retrieval



Training Data





Matching Sketches to Images

Classical Approach
shape matching

training data
(relatively cheap)

image

edge map

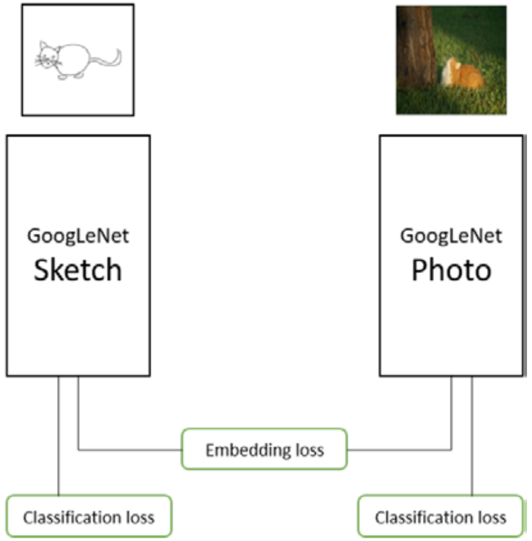
sketch

alignment

no training

Modern Approach
end-to-end deep learning

training data
(very expensive)



+ category + similarity
- man-years of annotation
- very difficult to train

Ours
deep shape matching

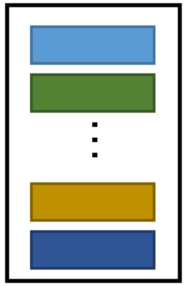
training data

image

edge map

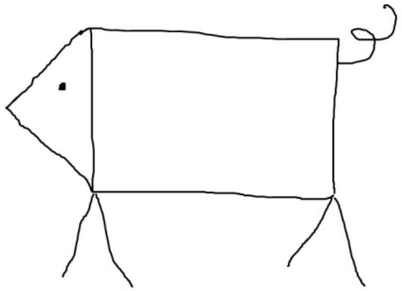
sketch

training data



shape information only
simple cost & training

Category Retrieval



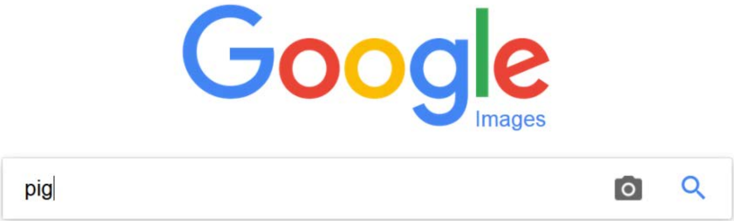
pig

Query

Result

Shape based retrieval cannot do that ☹

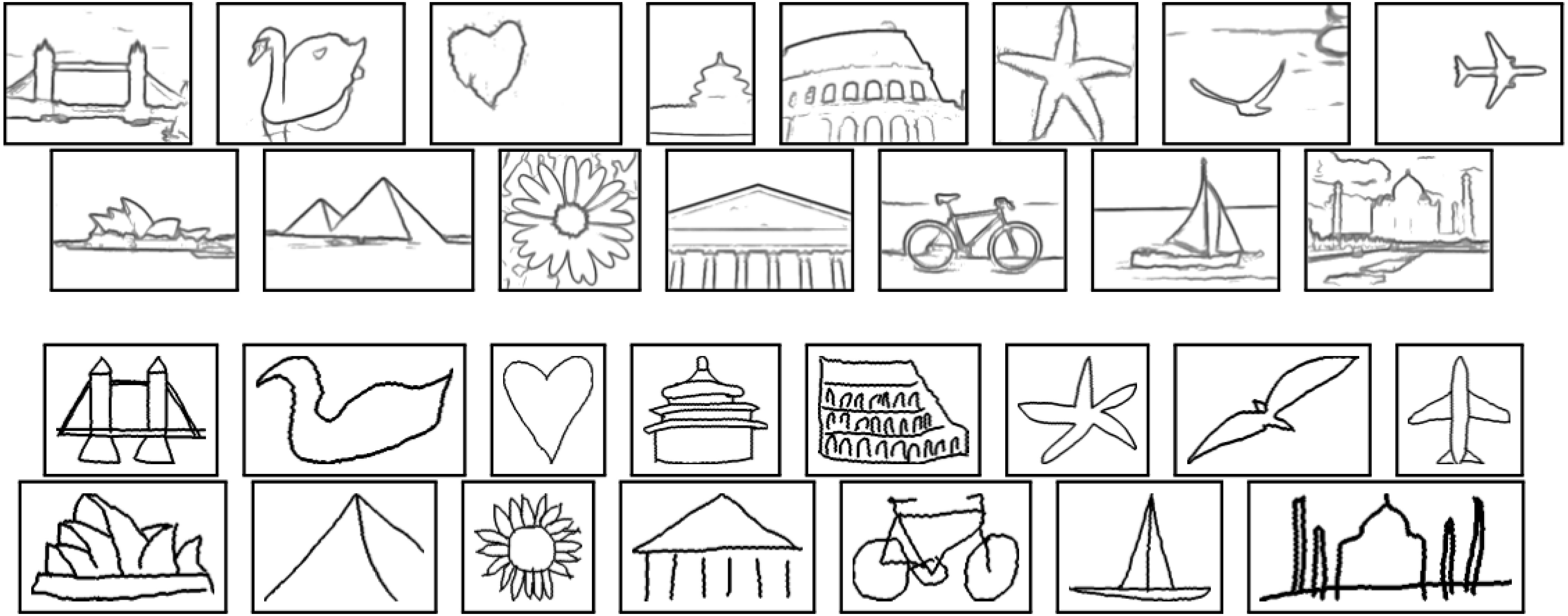
Category Retrieval



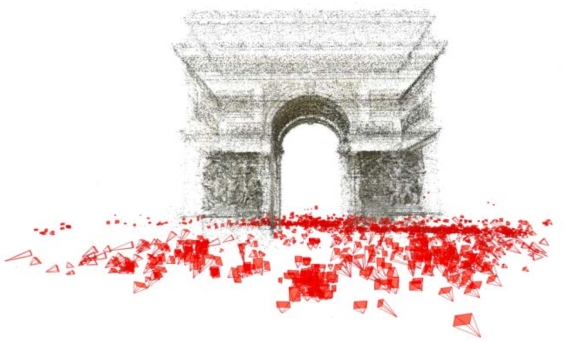
Result

Standard image search can do that for years already

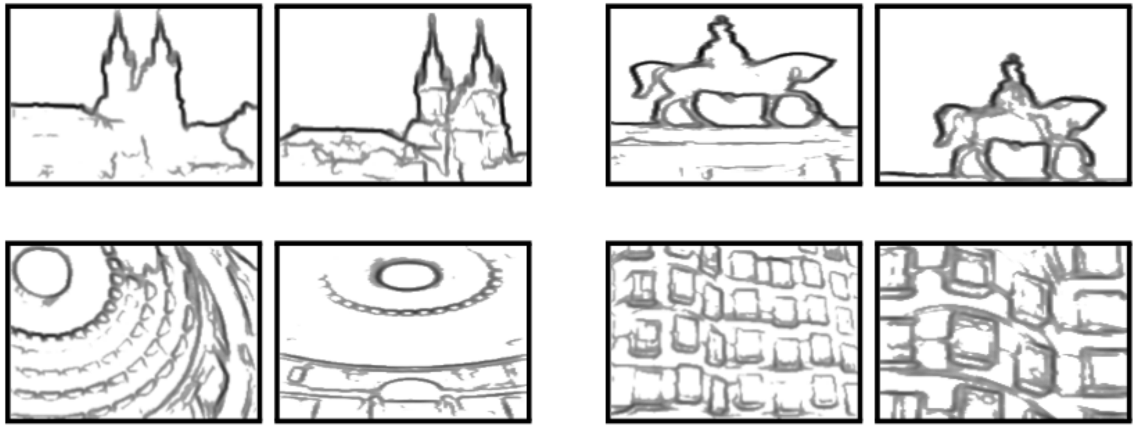
Edge-maps vs Sketches



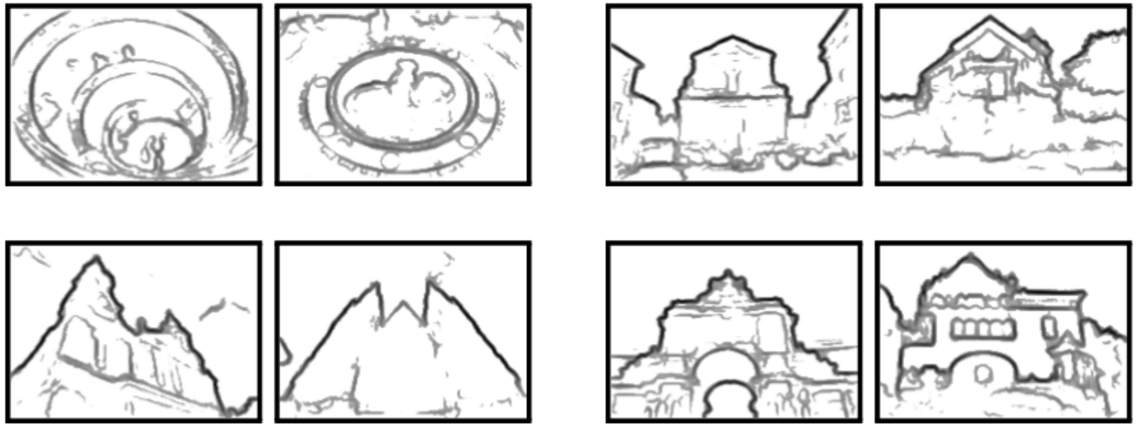
Training without a Single Sketch



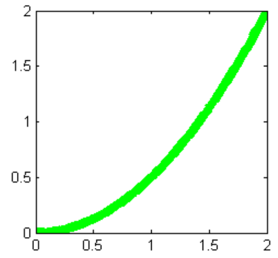
Positive (from geometrically verified images)

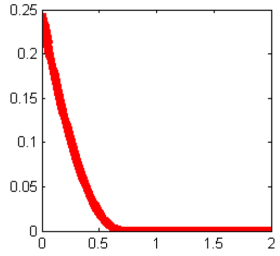


Negative (similar edge maps of different landmarks)

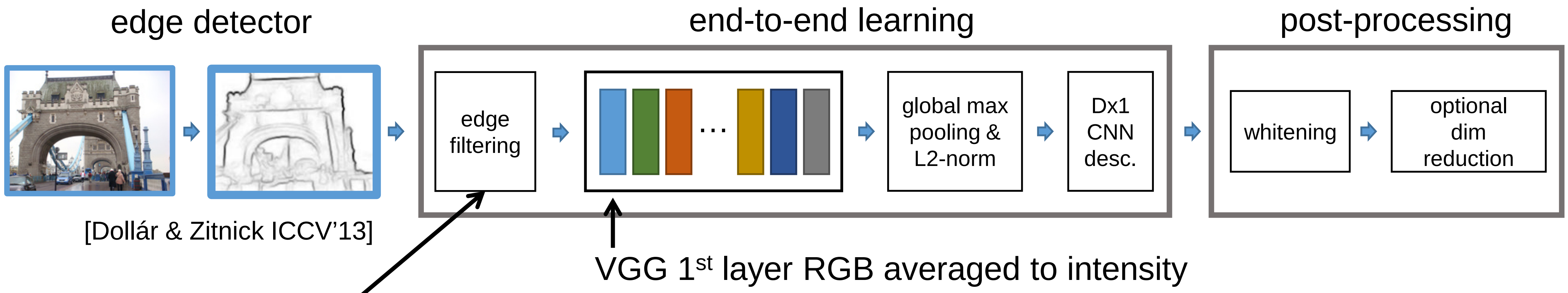


CNN Siamese learning
contrastive loss



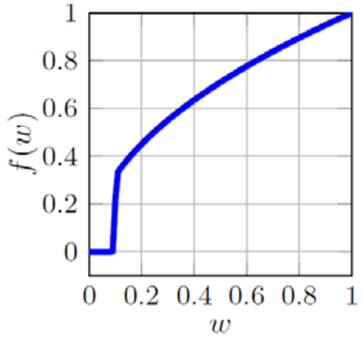


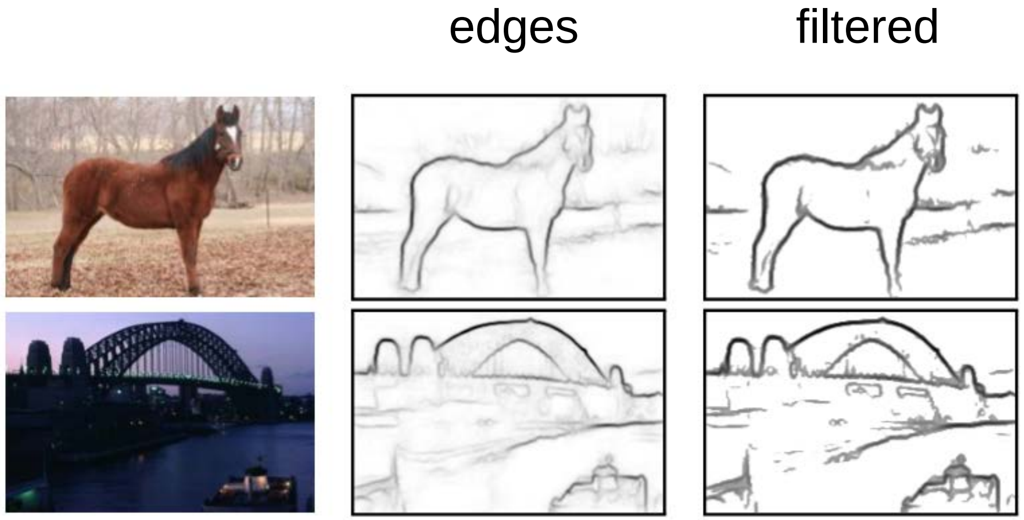
EdgeMAC Architecture



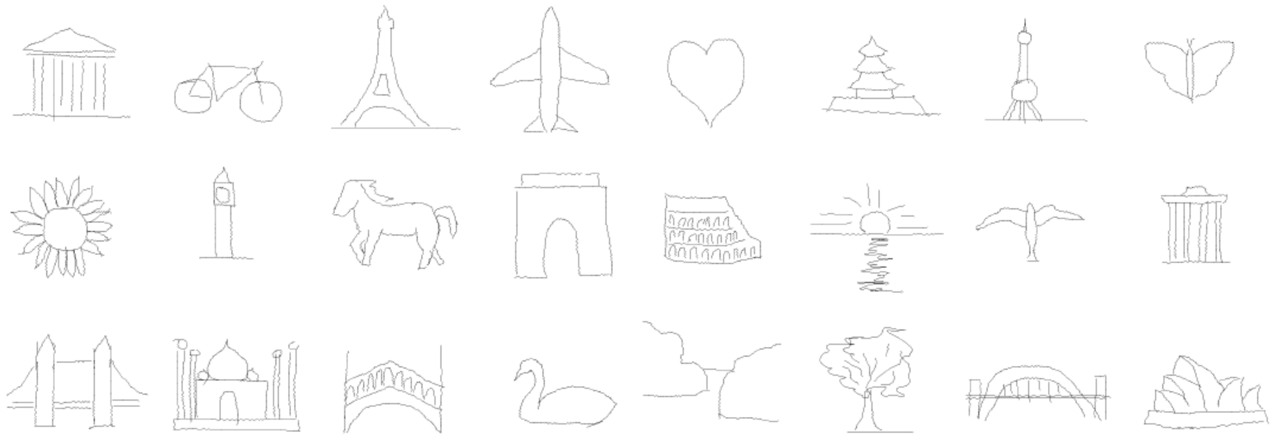
edge filtering layer

$$f(w) = \frac{w^p}{1 + e^{\beta(\tau - w)}}$$





Results on Flickr 15k



[21] Hu & Collomosse: **A performance evaluation of gradient field hog descriptor for sketch based image retrieval.** CVIU'13

Method	Dim	mAP
Hand-crafted methods		
GF-HOG [21]	n/a	12.2
S-HELO [37]	1296	12.4
HLR+S+C+R [51]	n/a	17.1
GF-HOG extended [6]	n/a	18.2
PerceptualEdge [32]	3780	18.4
LKS [38]	1350	24.5
AFM [47]	243	30.4
CNN-based methods		
Sketch-a-Net+EdgeBox [5]	5120	27.0
Siamese network [33]	64	19.5
Shoes network [53]†	256	29.9
Chairs network [53]†	256	29.8
Sketchy network [39]†	1024	34.0
Quadruplet network [41]	1024	32.2
Triplet no-share network [7]	128	**36.2**
⋆ EdgeMAC	512	**46.3**
Re-ranking methods		
AFM+QE [47]	755	**57.9**
Sketch-a-Net+EdgeBox+GraphQE [5]	n/a	32.3
⋆ EdgeMAC+Diffusion	n/a	**68.9**

Radenovic, Tolias, Chum: **Generic Sketch-Based Retrieval Learned without Drawing a Single Sketch**, arXiv 2017

Results on Shoes, Chairs and Handbags

Fine-grained recognition of shoes / chairs

[53] Q. Yu et al.: **Sketch me that shoe**. *CVPR'16.*

Image from https://www.eecs.qmul.ac.uk/~qian/Project_cvpr16.html

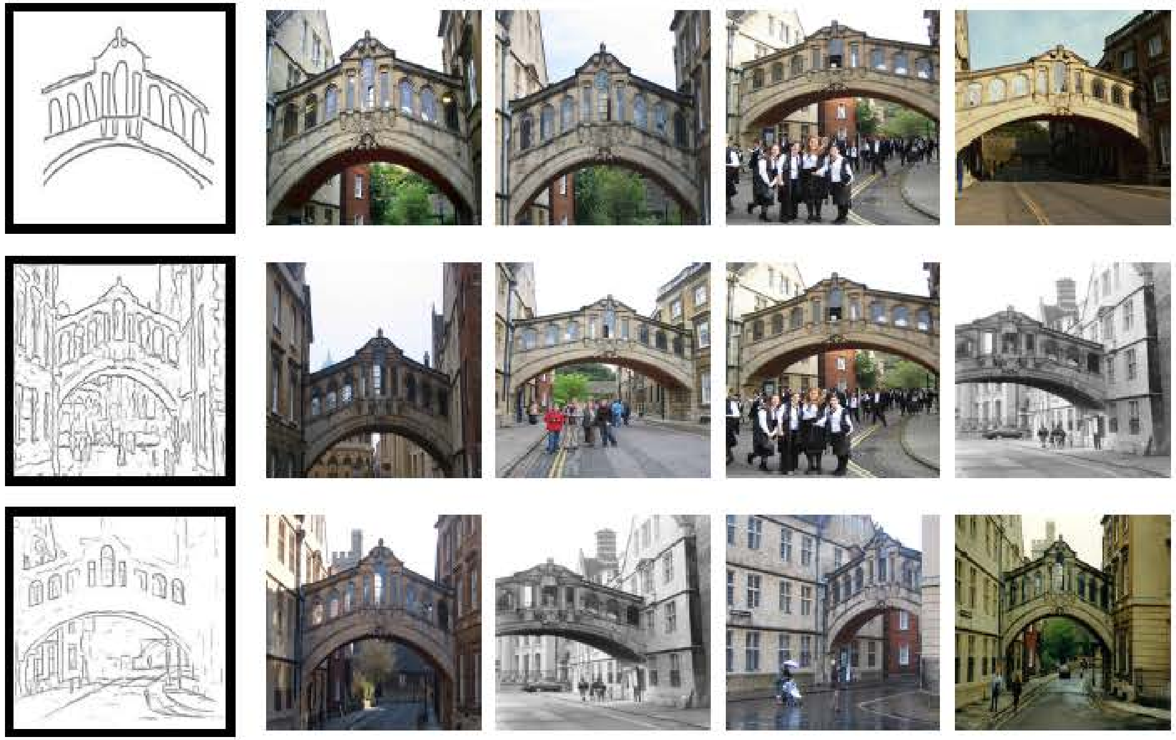
Results on Shoes, Chairs and Handbags

Method	Dim	Shoes		Chairs		Handbags	
		acc.@1	acc.@10	acc.@1	acc.@10	acc.@1	acc.@10
BoW-HOG + rankSVM [22]	500	17.4	67.8	28.9	67.0	2.4	10.7
Dense-HOG + rankSVM [22]	200K	24.4	65.2	52.6	93.8	15.5	40.5
Sketch-a-Net + rankSVM [22]	512	20.0	62.6	47.4	82.5	9.5	44.1
CCA-3V-HOG + PCA [18]	n/a	15.8	63.2	53.2	90.3	–	–
Shoes net [22]†	256	52.2	**92.2**	65.0	92.8	23.2	59.5
Chairs net [22]†	256	30.4	75.7	72.2	**99.0**	26.2	58.3
Handbags net [32]	256	–	–	–	–	39.9	82.1
Shoes net + CFF + HOLEF [32]	512	**61.7**	**94.8**	–	–	–	–
Chairs net + CFF + HOLEF [32]	512	–	–	**81.4**	95.9	–	–
Handbags net + CFF + HOLEF [32]	512	–	–	–	–	**49.4**	**82.7**
⋆ EdgeMAC	512	40.0	76.5	**85.6**	95.9	35.1	70.8
⋆ EdgeMAC + whitening	512	**54.8**	**92.2**	**85.6**	**97.9**	**51.2**	**85.7**

Beyond sketches

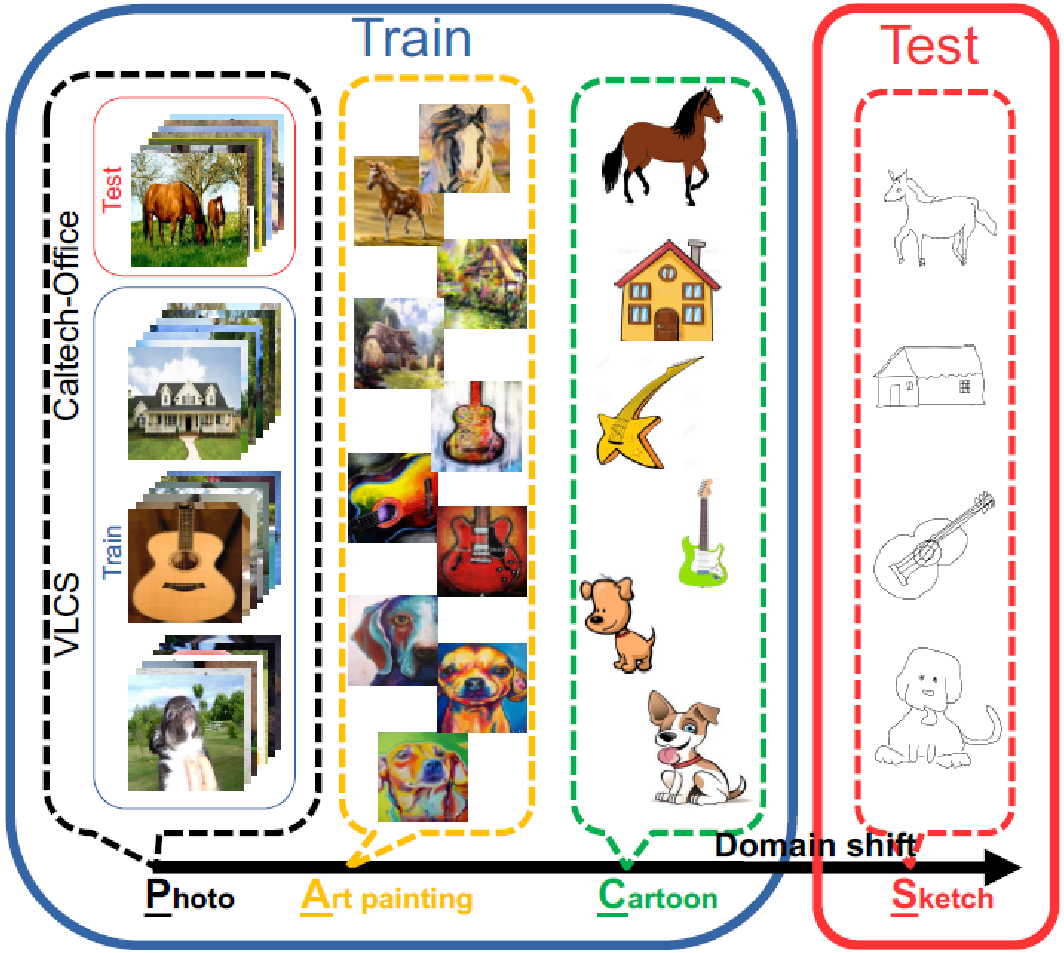
Image-based

Edge-based

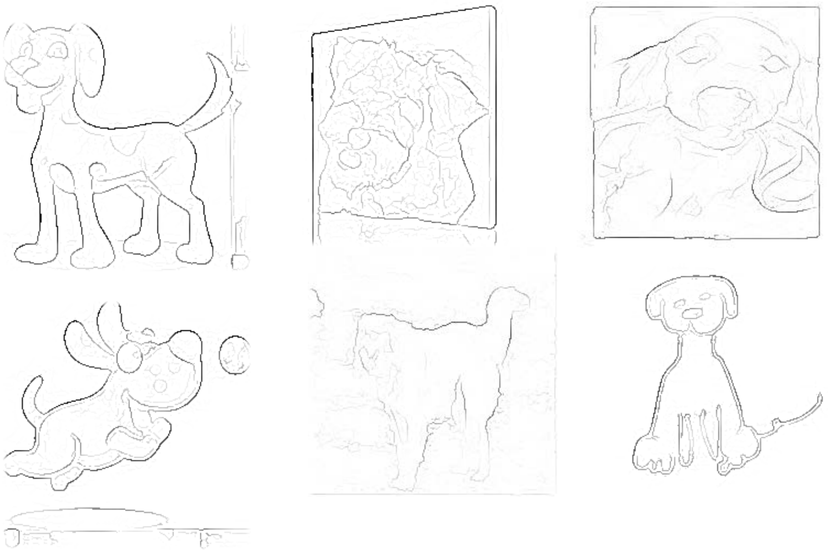


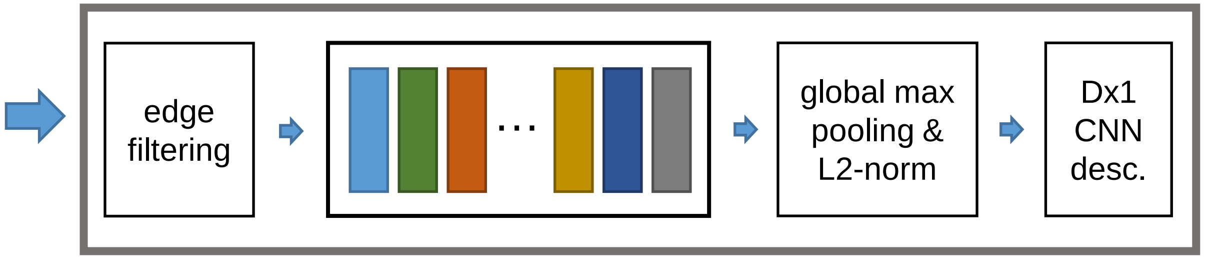
Shape matching for domain generalization

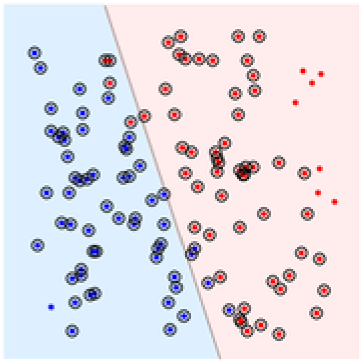
Domain generalization



Domain generalization via shape matching







Linear classifier on edgeMAC descriptors

Results on domain generalization

	Pre-trained (RGB)				Siamese [2] (RGB)				Ours (edge map)				Pre-trained+Ours			
Test →	A	C	P	S	A	C	P	S	A	C	P	S	A	C	P	S
Train A	N/A	59.2	95.0	33.1	N/A	59.5	86.3	42.9	N/A	55.9	61.2	65.6	N/A	61.6	94.9	38.4
Train C	71.7	N/A	86.8	37.0	61.0	N/A	77.0	51.6	45.2	N/A	57.3	74.8	69.3	N/A	85.0	55.3
Train P	72.5	33.3	N/A	24.8	66.0	38.0	N/A	31.9	45.4	42.3	N/A	46.3	73.3	34.0	N/A	27.61
Train S	31.9	49.5	42.5	N/A	38.7	49.3	44.4	N/A	34.8	63.0	43.3	N/A	33.7	59.3	43.4	N/A
Train 3	78.0	68.0	94.4	47.1	71.5	64.3	85.1	56.0	53.8	67.9	64.5	74.7	80.0	68.7	93.7	62.7
Mean 3	71.9				69.2				65.2				76.2			

A: Artwork C: Cartoon P: Photo S: Sketch

Metric Learning Without Labels

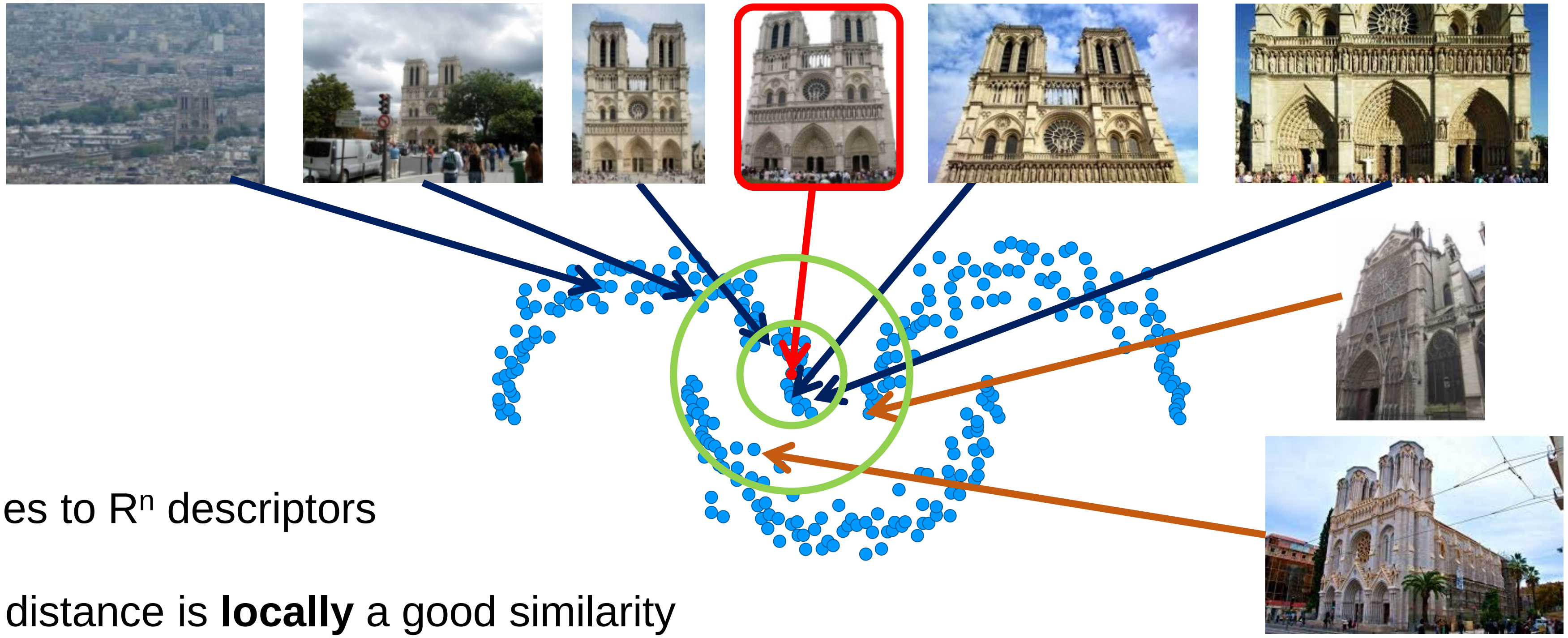
Ahmet Iscen

Giorgos Tolias

Yannis Avrithis

Teddy Furon

Euclidean & manifold distance



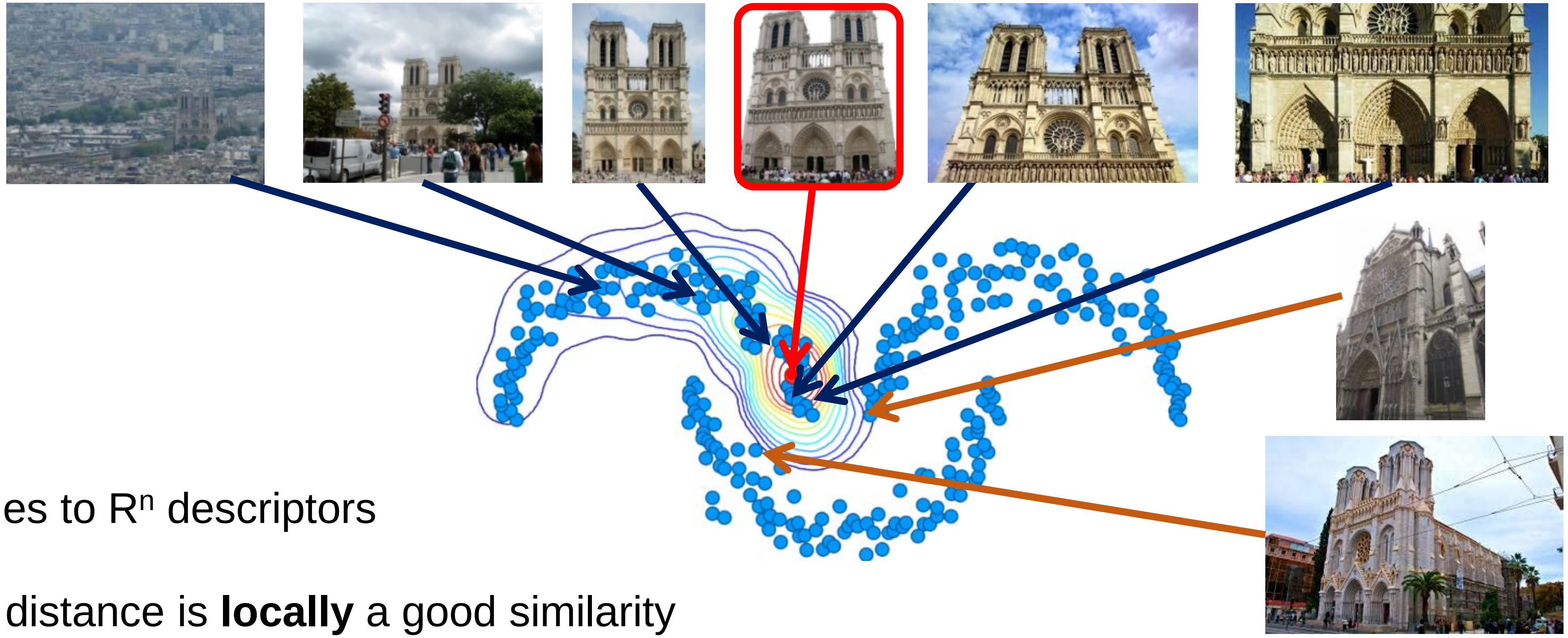
Mapping: Images to R^n descriptors

The Euclidean distance is **locally** a good similarity measure

Related images lie on non-linear manifolds

Iscen, Tolias, Avrithis, Furon, Chum, Efficient Diffusion on Region Manifolds: Recovering Small Objects with Compact CNN Representations, CVPR'17

Euclidean & manifold distance



<u>Mapping</u>: Images to R^n descriptors

The Euclidean distance is **locally** a good similarity measure

Related images lie on non-linear manifolds

Iscen, Tolias, Avrithis, Furon, Chum, Efficient Diffusion on Region Manifolds: Recovering Small Objects with Compact CNN Representations, CVPR’17

Diffusion

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

Normalized (sparse) affinity matrix

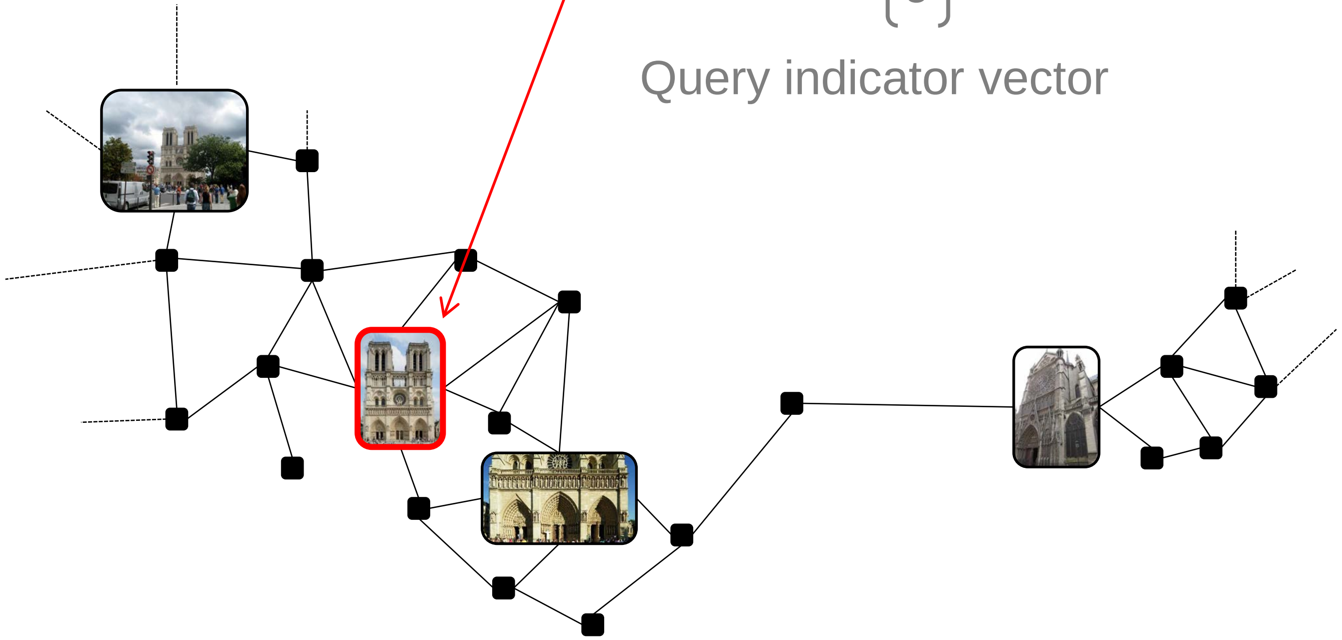
Vector of similarities to the query

$$\begin{bmatrix} 0 \\ 0 \\ \vdots \\ 1 \\ \vdots \\ 0 \\ 0 \end{bmatrix}$$

Query indicator vector

$$0 < \alpha < 1$$

Random walk
implicitly considers all paths
(visual proof)



k-Nearest Neighbour graph

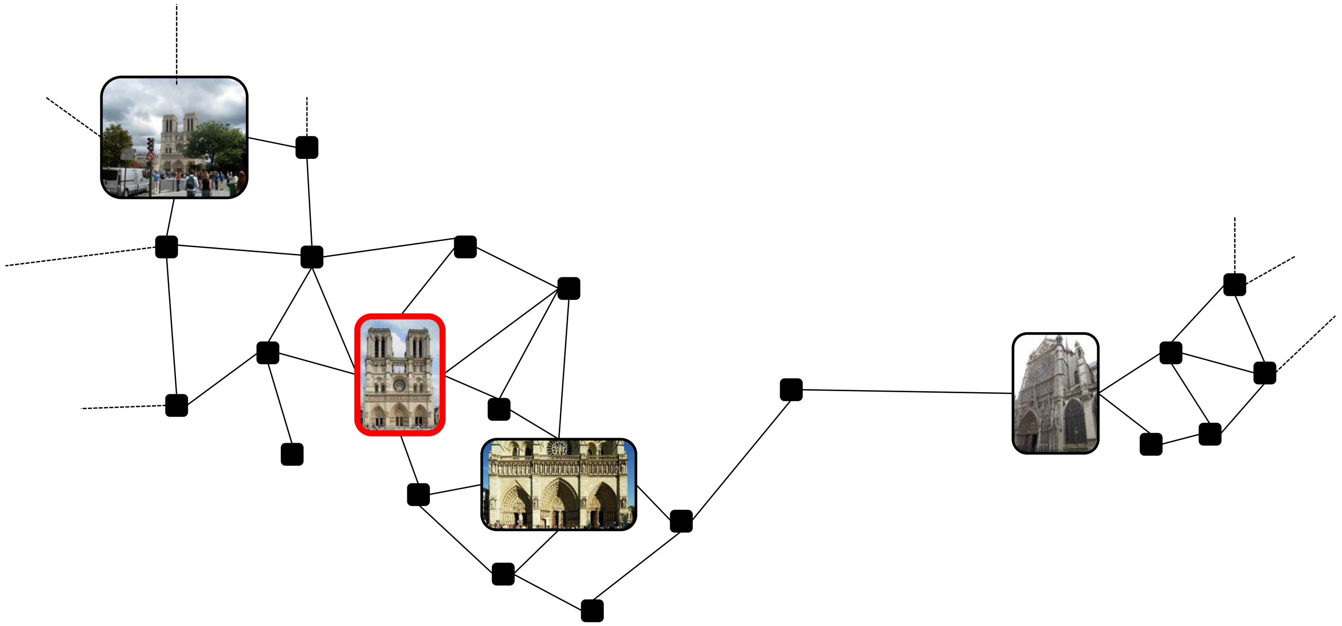
Diffusion

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

Closed form: $$\mathbf{f}^\star = \boxed{\mathcal{L}_\alpha^{-1}}\mathbf{y} \quad \text{where} \quad \boxed{\mathcal{L}_\alpha} = \frac{I_n - \alpha S}{1-\alpha}$$

Large, non-sparse

Large, sparse



k-Nearest Neighbour graph

Contributions on Diffusion for Retrieval

Iterative: $$\mathbf{f}^t = \alpha S \mathbf{f}^{t-1} + (1-\alpha)\mathbf{y}$$ Jacobi solver

Closed form: $$\mathbf{f}^\star = \boxed{\mathcal{L}_\alpha^{-1}}\mathbf{y}$$ Intractable

- $\boxed{\mathcal{L}_\alpha}\mathbf{f}^\star = \mathbf{y}$ System of linear equations, Conjugate Gradients [CVPR 2017]
- Generalization to novel queries (not part of the dataset)
- Diffusion can be efficiently applied to image parts
  - Significant impact on CNN-based retrieval of small object
- $\mathbf{f}^\star = \boxed{\mathcal{L}_\alpha^{-1}}\mathbf{y} \approx \boxed{U}\Lambda'\boxed{U}^\top\mathbf{y}$ Low-rank approximation
  - Two orders of magnitude faster online diffusion

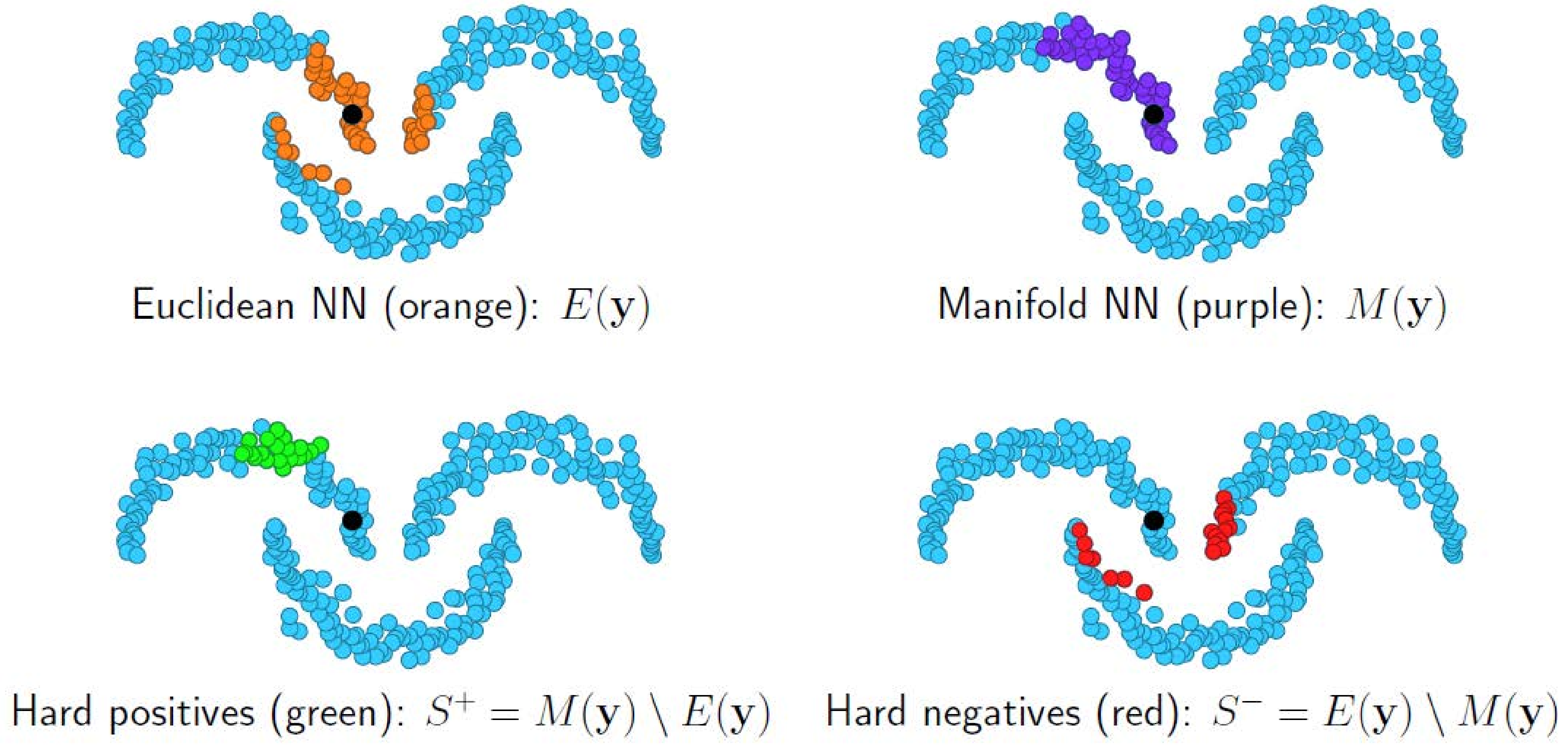
Small, non-sparse

[CVPR 2018]
[ACCV 2018]

Euclidean vs Manifold Distance

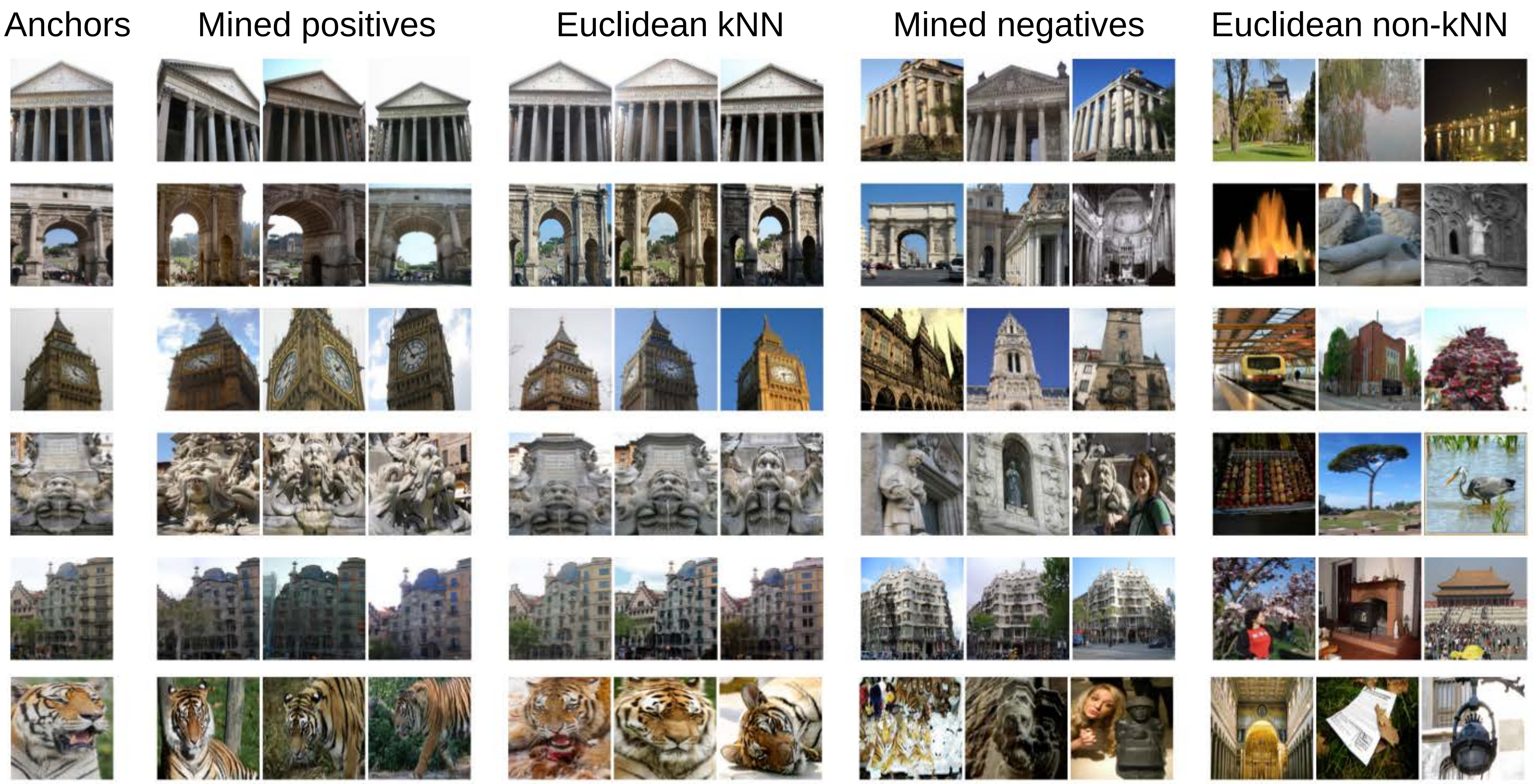
Diffusion-guided to sample hard negatives and positives

- Avoid computationally expensive SfM models

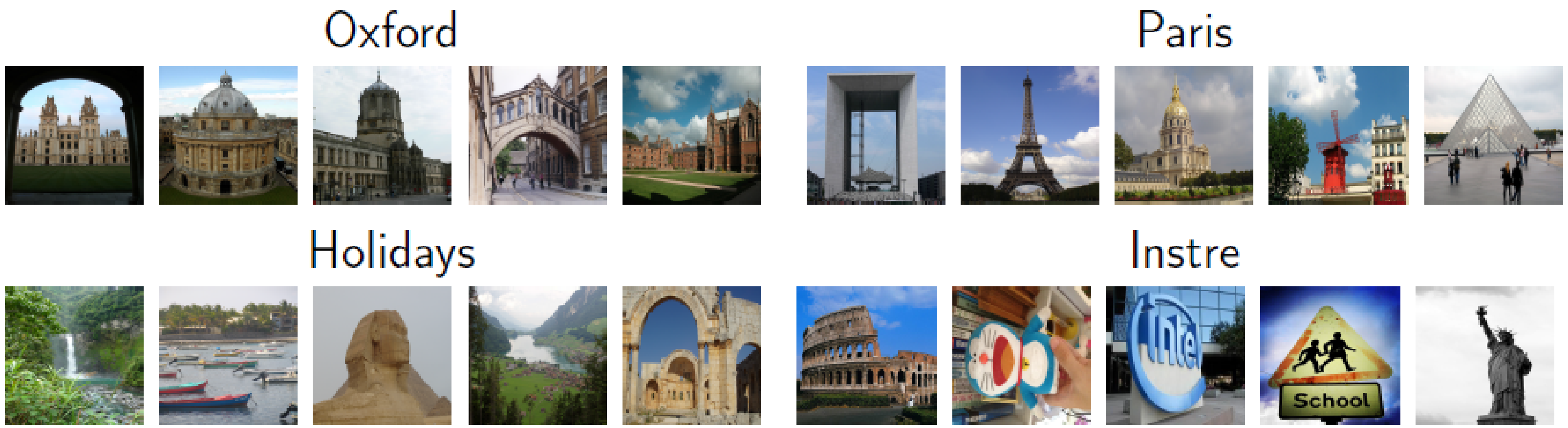


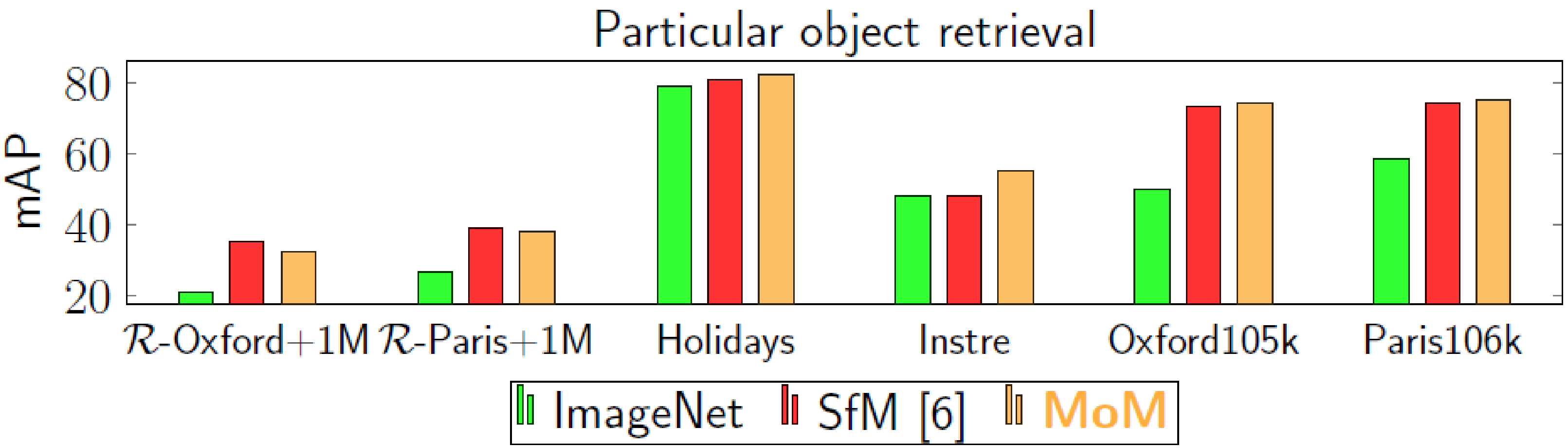
A. Iscen, G. Tolias , Y. Avrithis, O. Chum, Mining on Manifolds: Metric Learning without Labels, In CVPR 2018

Mining of training samples

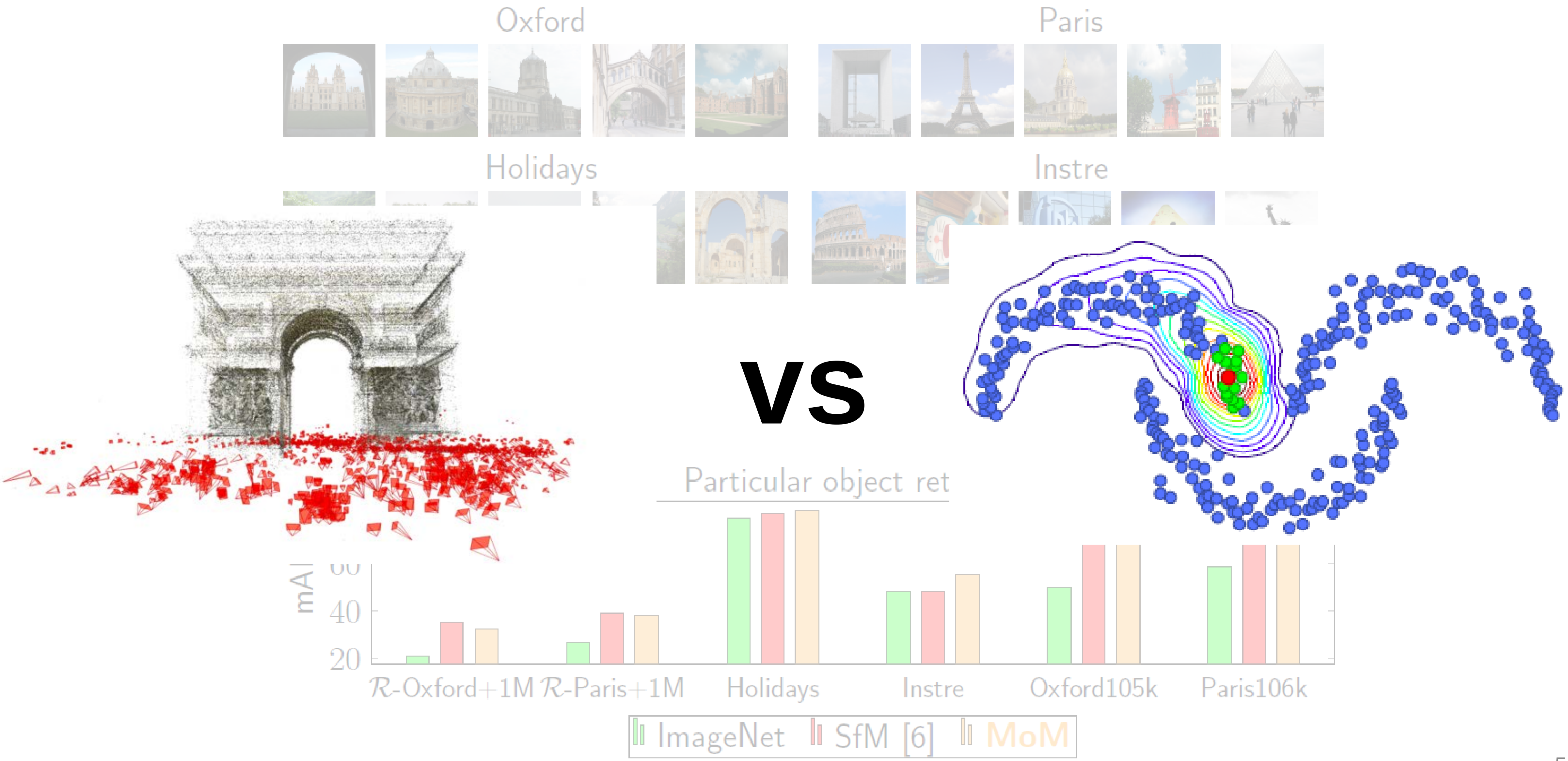


Experiments on instance search

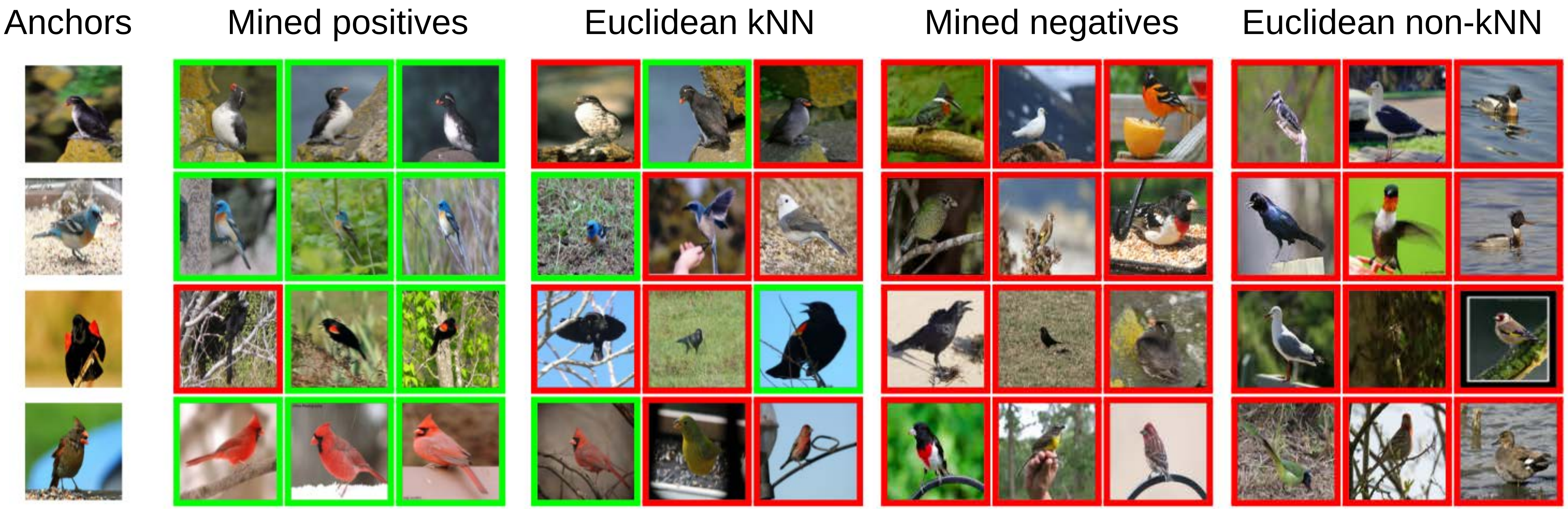




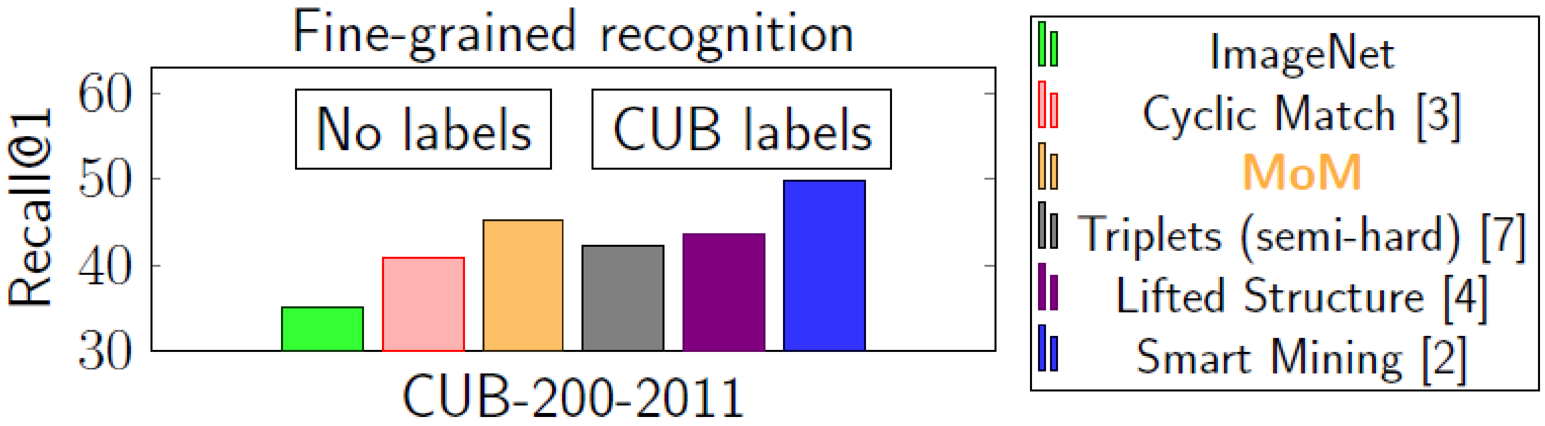
Experiments on instance search



Mining of training samples



Experiments on fine-grained recognition



Online code and data

Siamese training code and training data

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

- Image retrieval (ECCV 2016)
- Matlab package using MatConvNet
- Python package using PyTorch
- Sketch based image retrieval (ECCV 2018)
- Matlab package using MatConvNet

Region manifold search (CVPR 2017)

https://github.com/ahmetius/diffusion-retrieval

- Matlab package

Conclusions

BOW combined SfM is a good teacher

- no human annotation needed for CNN image retrieval
- CNN outperforms its teacher on standard benchmarks
- BOW still better for certain tasks

- no human annotation needed for CNN sketch based retrieval
- generic CNN shape retrieval performs well
  - standard and fine-grained sketch based retrieval
  - significant appearance changes, domain generalization

Mining on Manifolds

- fine tuning CNNs without supervision
- using diffusion to compute manifold distance

Thank you.