Visual Retrieval with Compact Image Representations

Filip Radenović

PhD Thesis Defence

Supervisor: Ondřej Chum

Visual Retrieval

Query
Image

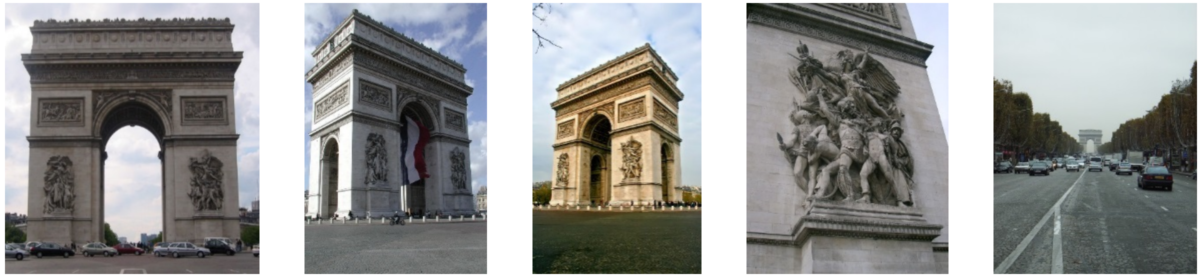
**Image
Retrieval
System**

Large Internet
photo collection

Retrieved Images

Addressed Challenges

Viewpoint and/or scale change



Visually similar but different

Illumination change

Different image modalities

Occlusion

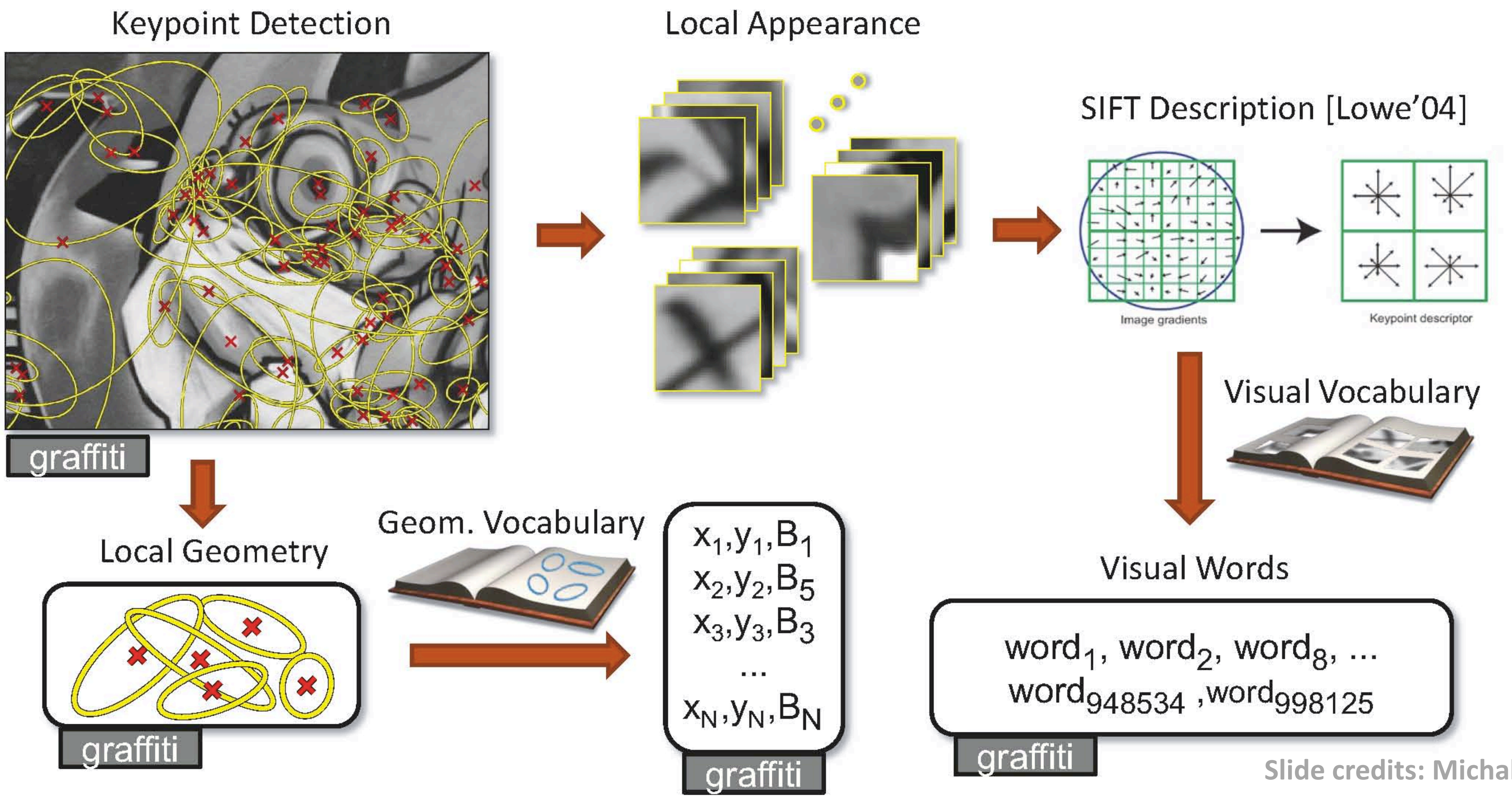
Billions of images

- Memory requirement
- Processing time
- Search time

Improving Bag-of-Words-Based Compact Image Retrieval

F. Radenovic, H Jegou, O. Chum. Multiple Measurements and Joint Dimensionality Reduction for Large Scale Image Search with Short Vectors. ICMR, 2015.

Bag-of-Words (BoW) approach



Slide credits: Michal Perdoch

Sivic, Zisserman: Video Google, ICCV 2003
Philbin, Chum, Isard, Sivic, Zisserman: Object retrieval with large vocabularies and fast spatial matching, CVPR 2007

PCA dimensionality reduction

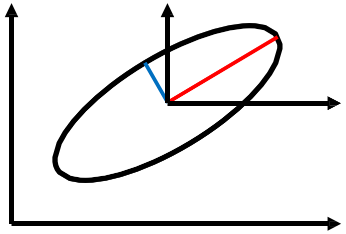
High dimensional sparse
BOW image representation

*Search is done using inverted files

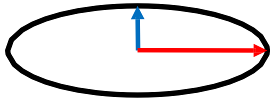
PCA

128 dimensional dense
image representation

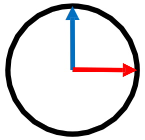
*Search is done using (approximate) nearest-neighbors



- Centering – emphasize negative evidence, higher importance of jointly missing visual words



- PCA rotation – decorrelating and allowing to remove least informative dimensions

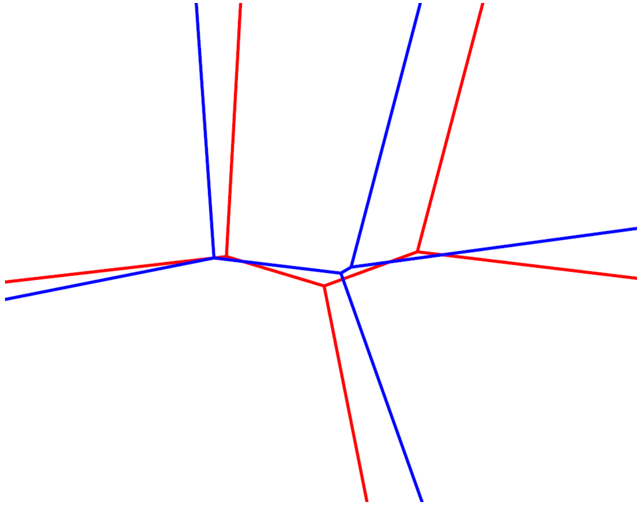


- Whitening – addresses over-counting (burstiness, co-occurence)

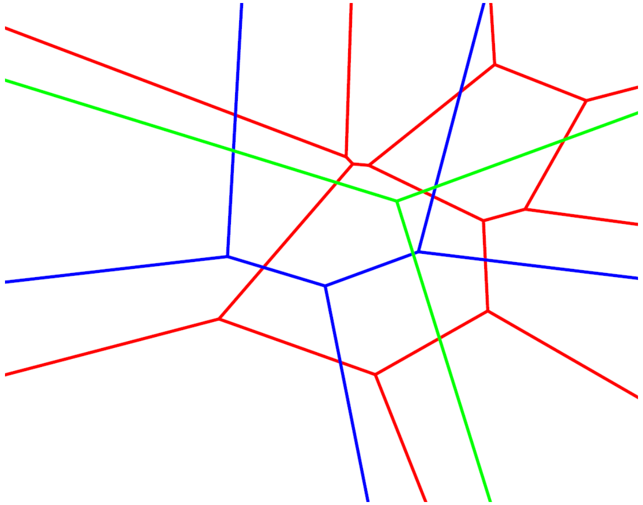
Jegou, Chum: Negative evidences and co-occurrences in image retrieval: the benefit of PCA and whitening, ECCV 2012

PCA reduction of multiple vocabularies

1. Multiple vocabularies are built using different k-means initializations
2. BOW vectors are concatenated
3. Concatenated BOW vectors are jointly PCA-reduced and whitened



Different vocabulary initializations



Different vocabulary sizes

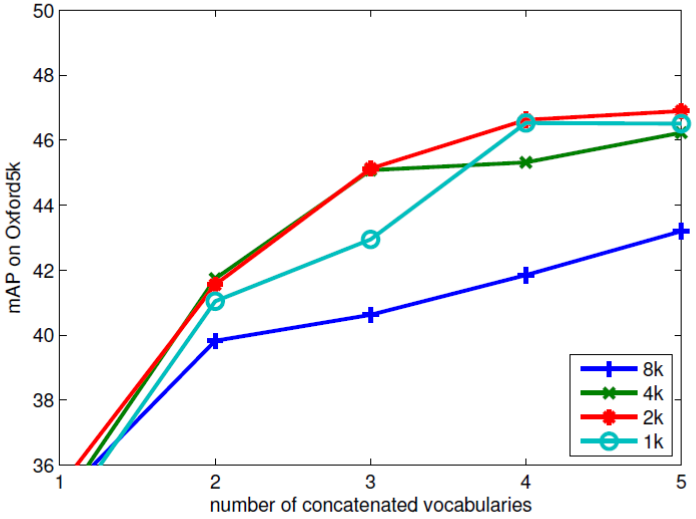
Jegou, Chum: Negative evidences and co-occurrences in image retrieval: the benefit of PCA and whitening, ECCV 2012

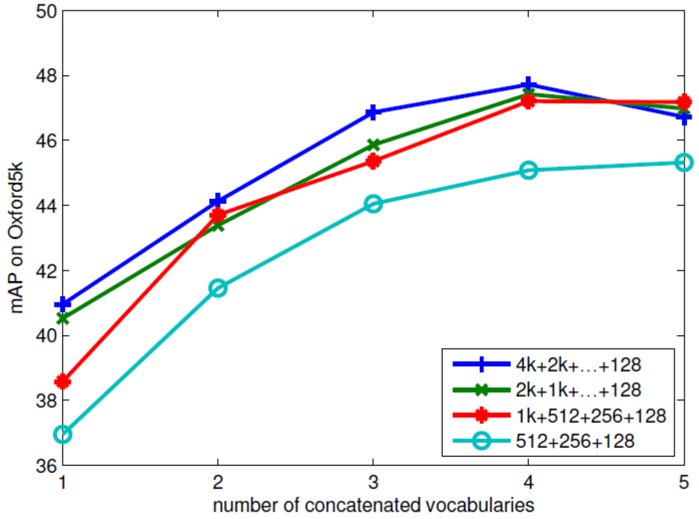
Multiple measurement regions

Construct vocabularies at multiple relative scales of the measurement regions:

$0.5 \times r$ $0.75 \times r$ $1 \times r$ $1.25 \times r$ $1.5 \times r$

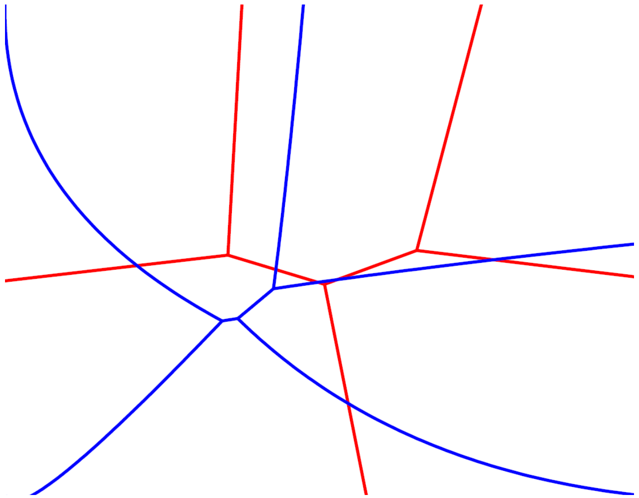
r = 3√3 – relative change in the measured area radius compared to detected area radius

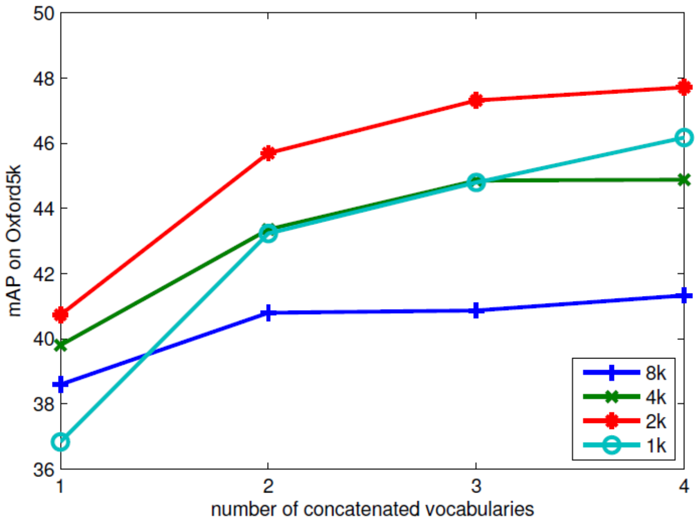


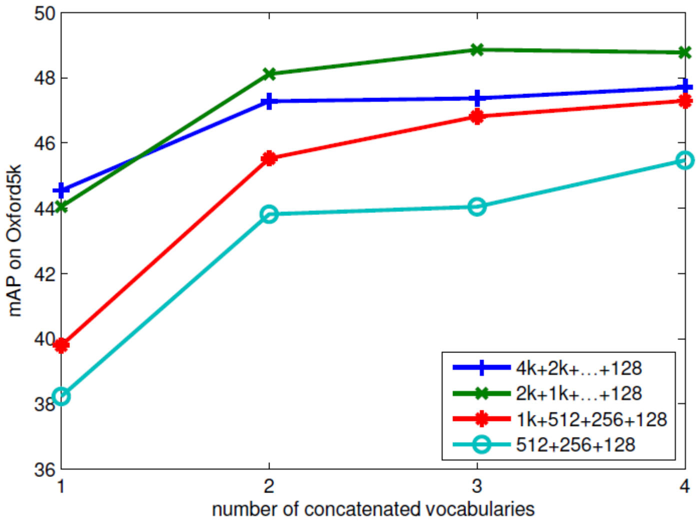


Multiple rooted SIFT descriptors

- Combine SIFT and SIFT with every component to the power of 0.4 ($SIFT^{0.4}$), 0.5 ($SIFT^{0.5}$), 0.6 ($SIFT^{0.6}$) to create four different vocabularies
- SIFT descriptors + Euclidian = hyperplanes
- RootSIFTs + Euclidian = curved hypersurfaces in SIFT space







Training Convolutional Neural Networks for Image Retrieval

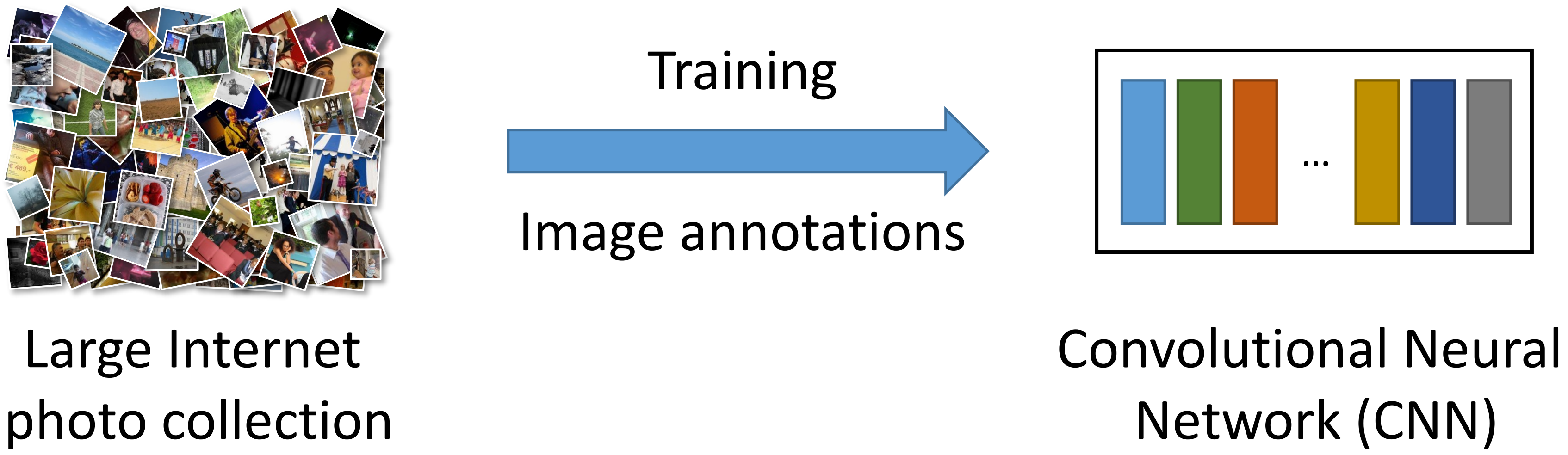
J. L. Schonberger, F. Radenovic, O. Chum, J. Frahm. From Single Image Query to Detailed 3D Reconstruction. CVPR, 2015.

F. Radenovic, J. L. Schonberger, D. Ji, J. Frahm, O. Chum, J. Matas. From Dusk till Dawn: Modeling in the Dark. CVPR, 2016.

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

F. Radenovic, G. Tolias, O. Chum. Fine-tuning CNN Image Retrieval with No Human Annotation. TPAMI, 2018.

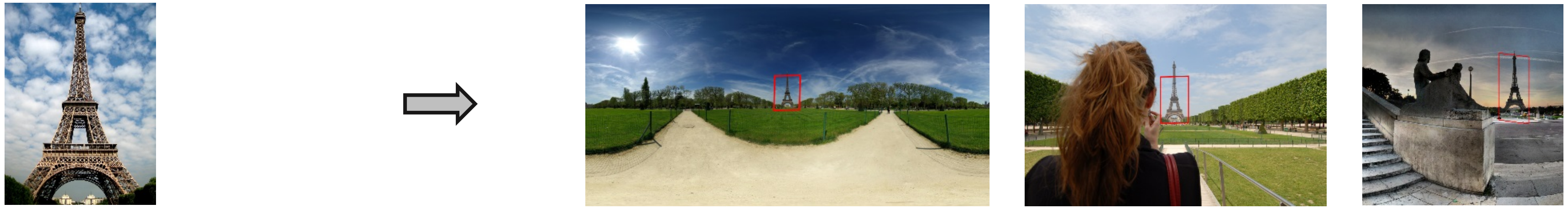
Training dataset



Retrieval-Structure-from-Motion pipeline

Visually most similar

Zoom-in / details



Zoom-out

Sideways right

Retrieval-Structure-from-Motion pipeline

Camera Orientation Known
Number of Inliers Known

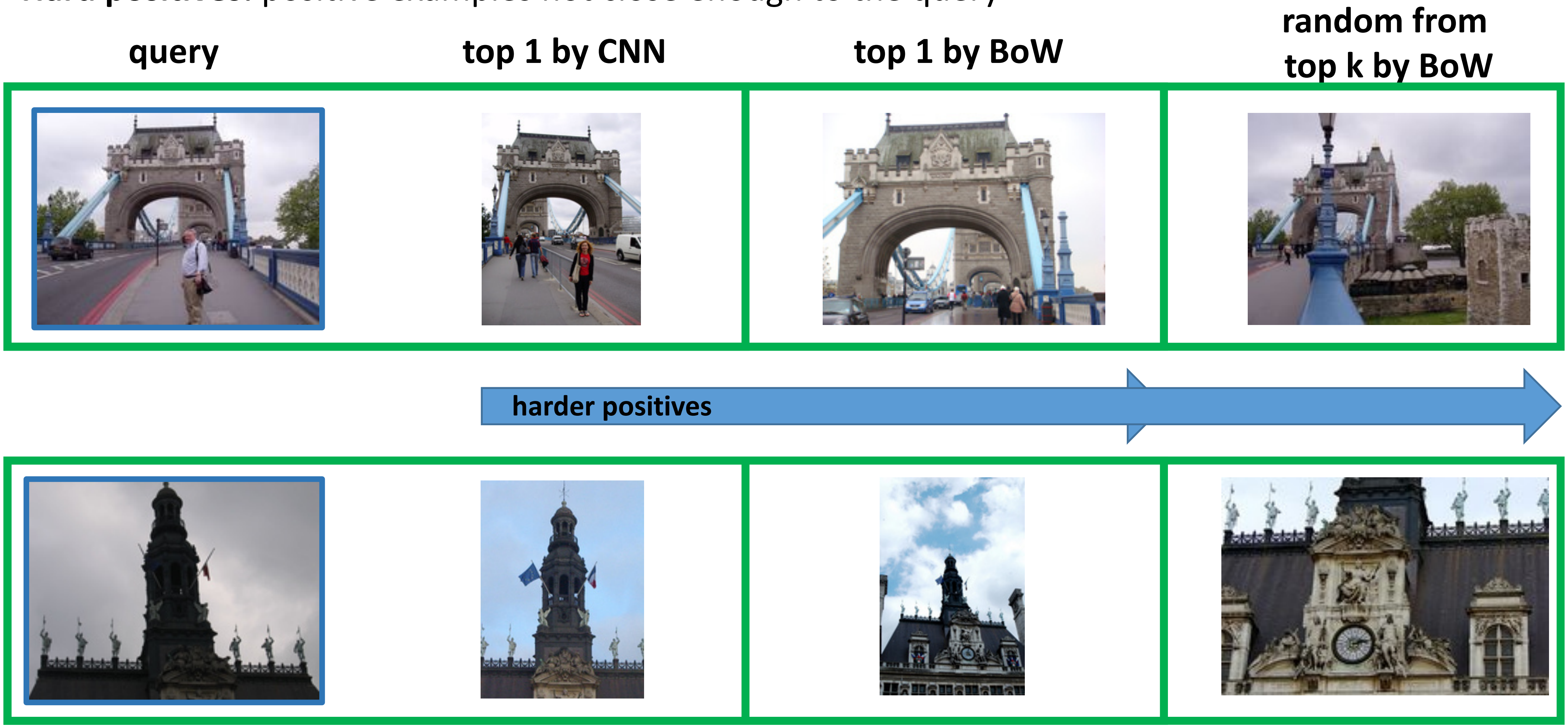
7.4M images → 713 training 3D models

Hard negative examples

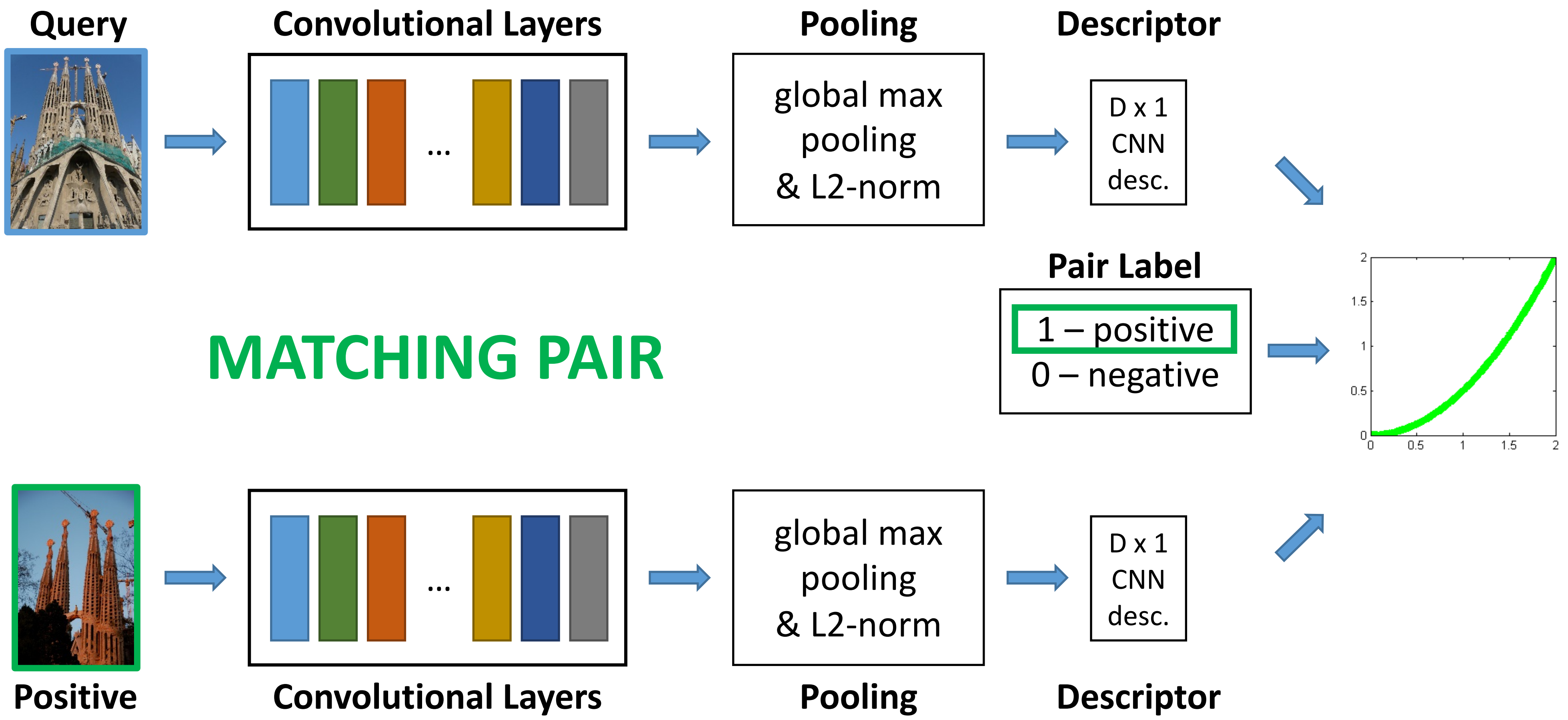
Negative examples: images from different 3D models than the query
Hard negatives: closest negative examples to the query
Only hard negatives: as good as using all negatives, but faster

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

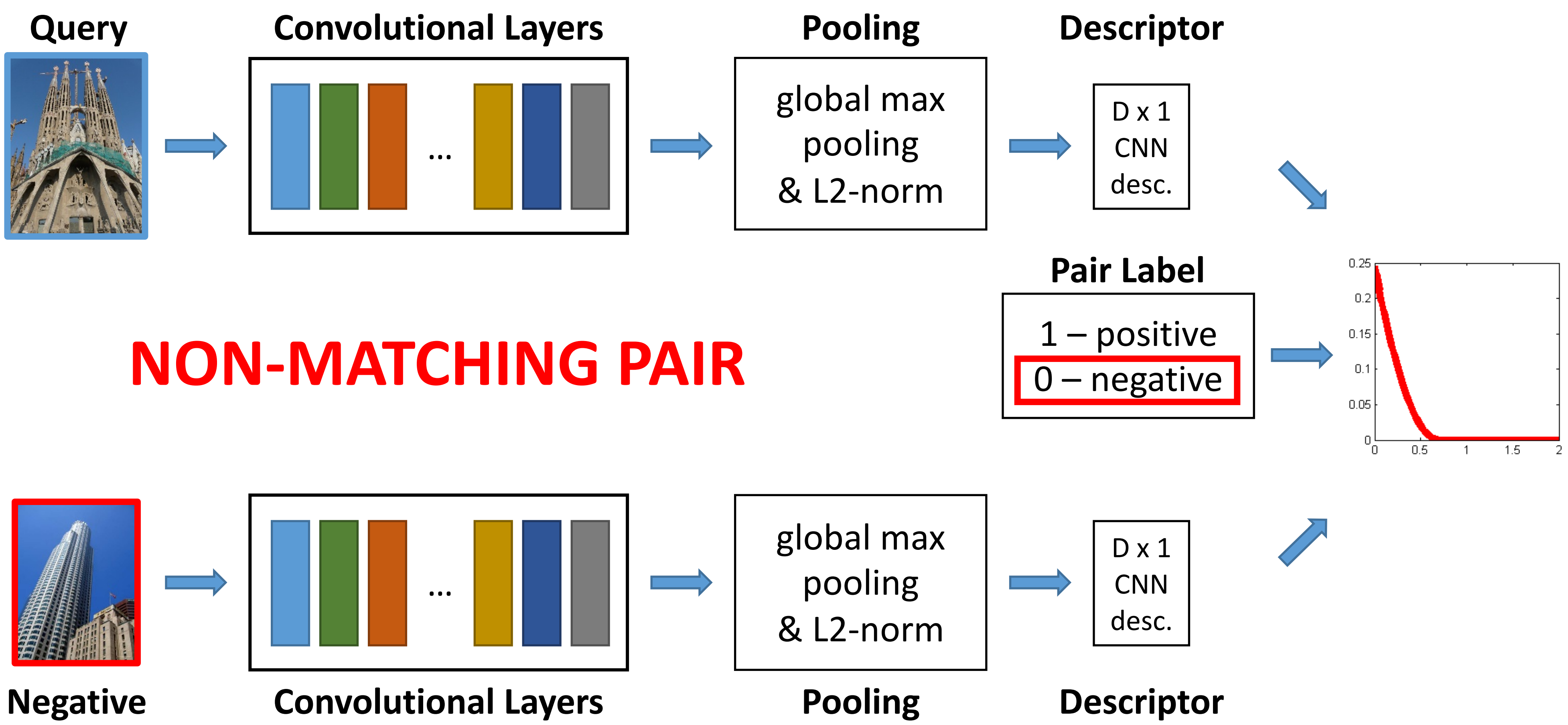
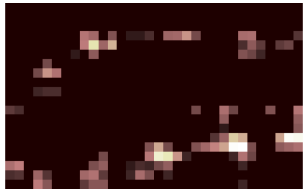


Image representation

Input image

$\mathcal{X}_1$



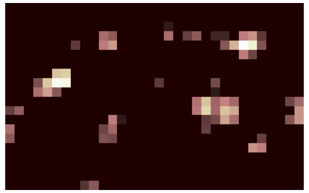
conv$_5$ filter 1

$\mathcal{X}_2$

conv$_5$ filter 2

....

$\mathcal{X}_k$



conv$_5$ filter k

....

$\mathcal{X}_K$

conv$_5$ filter K

Image descriptor: $\boldsymbol{f} = [f_1 \dots f_k \dots f_K]$

Max pooling (MAC): $f_k = \max_{x \in \mathcal{X}_k} x$

Sum pooling (SpOC): $f_k = \frac{1}{|\mathcal{X}_k|} \sum_{x \in \mathcal{X}_k} x$

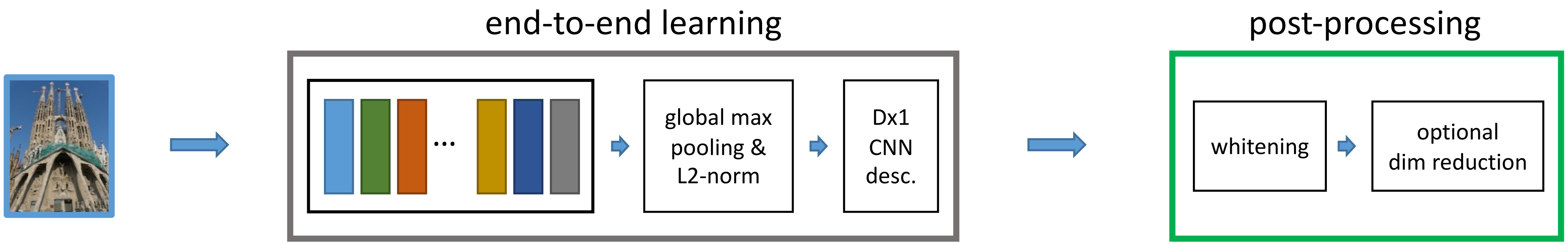
Generalized-mean pooling (GeM):

$$f_k = \left(\frac{1}{|\mathcal{X}_k|} \sum_{x \in \mathcal{X}_k} x^p \right)^{\frac{1}{p}}$$

$p \to \infty$ MAC

$p = 1$ SPoC

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

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!!!

Image Retrieval: State of the Art Evaluation

F. Radenovic, A. Iscen, G. Tolias, Y. Avrithis, O. Chum. Revisiting Oxford and Paris: Large-Scale Image Retrieval Benchmarking. CVPR, 2018.

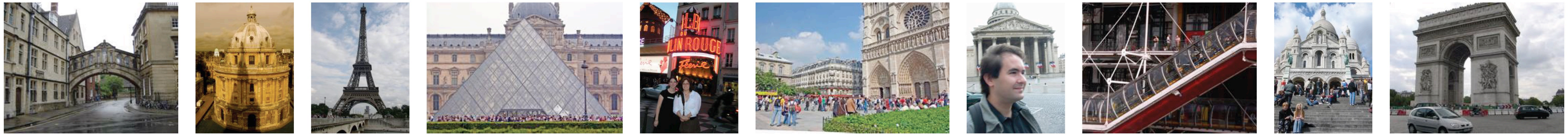
Revisiting Oxford and Paris: What was wrong?

- **Annotation errors:** skewed comparison of different methods

Original labeling mistakes: **Query (blue)** image and the associated database images that were originally marked as **negative (red)** or **positive (green)**.

- **Solved:** saturated performance, every challenging image labeled as *Junk*

- **Over-fitting:** small datasets, extension Oxford 100k (easy, false negatives)



Examples of false negative images in Oxford100k.

Revisiting Oxford and Paris: What is new?

- Errors in the annotation are fixed
- ***Labeling of all*** images is revisited
- New distractor dataset with 1 million images is created
- Images are chosen to be challenging for these two benchmarks
- New set of 15 queries per benchmark is added
- New set of evaluation protocols with increasing difficulty:
  Easy (E), Medium (M), and Hard (H)

State of the art evaluation

Time and Memory

Method	Memory (GB)	Time (sec) Extraction GPU	Time (sec) Extraction CPU	Time (sec) Search
HesAff–rSIFT–ASMK⋆	62.0	n/a + 0.06	1.08 + 2.35	0.98
HesAff–rSIFT–ASMK⋆+SP				2.00
DELF–ASMK⋆+SP	10.3	0.41 + 0.01	n/a + 0.54	0.52
A–[FT]–GeM	0.96	0.12	1.99	0.38
V–[FT]–GeM	1.92	0.23	31.11	0.56
R–[FT]–GeM	7.68	0.37	14.51	1.21

mAP Old vs New

Method	Oxf	$\mathcal{R}$Oxford E	$\mathcal{R}$Oxford M	$\mathcal{R}$Oxford H	Par	$\mathcal{R}$Paris E	$\mathcal{R}$Paris M	$\mathcal{R}$Paris H
HesAff–rSIFT–SMK⋆	78.1	74.1	59.4	35.4	74.6	80.6	59.0	31.2
R–[O]–R-MAC	78.3	74.2	49.8	18.5	90.9	89.9	74.0	52.1
R–[FT]–GeM	87.8	84.8	64.7	38.5	92.7	92.1	77.2	56.3
R–[FT]–GeM+DFS	90.0	86.5	69.8	40.5	95.3	93.9	88.9	78.5

State-of-the-art performance

Method	Medium $\mathcal{R}$Oxf+$\mathcal{R}$1M mAP	Medium $\mathcal{R}$Oxf+$\mathcal{R}$1M mP@10	Medium $\mathcal{R}$Par+$\mathcal{R}$1M mAP	Medium $\mathcal{R}$Par+$\mathcal{R}$1M mP@10	Hard $\mathcal{R}$Oxf+$\mathcal{R}$1M mAP	Hard $\mathcal{R}$Oxf+$\mathcal{R}$1M mP@10	Hard $\mathcal{R}$Par+$\mathcal{R}$1M mAP	Hard $\mathcal{R}$Par+$\mathcal{R}$1M mP@10
HesAff–rSIFT–VLAD	17.4	34.8	19.6	76.1	5.6	7.0	3.3	21.1
HesAff–rSIFT–SMK⋆+SP	38.1	67.1	34.5	89.3	17.7	30.3	11.0	49.1
HesAff–rSIFT–ASMK⋆+SP	46.8	79.6	42.3	95.3	26.9	45.3	16.8	65.3
DELF–ASMK⋆+SP	53.8	81.1	57.3	98.3	31.2	50.7	26.4	75.7
R – [O] –MAC	24.2	43.7	40.8	93.0	5.7	14.4	18.2	67.7
R – [O] –SPoC	21.5	40.4	41.6	92.0	2.8	5.6	15.3	54.4
R – [O] –CroW	21.2	39.4	42.7	92.9	3.3	9.3	16.3	61.6
R – [O] –GeM	25.6	45.1	46.2	94.0	4.7	13.4	20.3	70.4
R – [O] –R-MAC	29.2	48.9	49.3	93.7	4.5	13.0	21.3	67.4
R – [FT] – GeM	45.2	71.7	52.3	95.3	19.9	34.9	24.7	73.3
R – [FT] – R-MAC	39.3	62.1	54.8	93.9	12.5	24.9	28.0	70.0
Query expansion (QE) and diffusion (DFS)								
HesAff–rSIFT–HQE	42.7	67.4	44.2	90.1	23.2	37.6	20.3	51.4
HesAff–rSIFT–HQE+SP	52.0	76.7	46.8	93.0	29.8	50.1	21.8	61.9
DELF–HQE+SP	60.6	79.7	65.2	96.1	37.9	56.1	35.8	69.1
R – [FT] – GeM+αQE	49.0	74.7	58.0	95.9	24.2	40.3	31.0	80.4
R – [FT] – GeM+DFS	61.5	77.1	84.9	95.9	33.1	48.2	71.6	93.7
R – [FT] – R-MAC+DFS	56.6	68.6	83.2	93.3	28.4	43.6	70.4	89.1
HesAff–rSIFT–ASMK⋆+SP $\rightarrow$ R–[FT]–GeM+DFS	74.3	87.9	85.9	97.1	48.7	65.9	73.2	96.6
HesAff–rSIFT–ASMK⋆+SP $\rightarrow$ R–[FT]–R-MAC+DFS	74.9	87.9	87.5	97.1	47.5	62.4	76.0	96.3
DELF–ASMK⋆+SP $\rightarrow$ R–[FT]–R-MAC+DFS	68.7	83.6	86.6	98.1	39.4	55.7	74.2	94.6

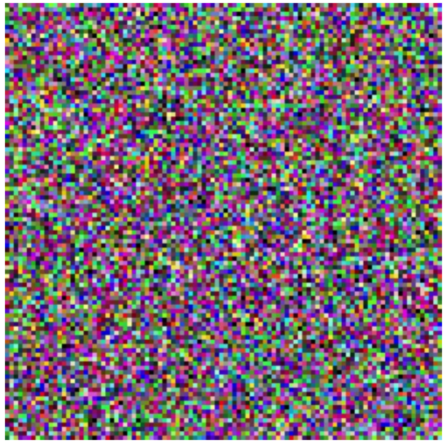
Targeted Mismatch Adversarial Attack to Conceal the Query Image

G. Tolias, F. Radenovic, O. Chum. Targeted Mismatch Adversarial Attack: Query with a Flower to Retrieve the Tower. ICCV, 2019.

Misclassification Adversarial Attack

“cat”

+ **ε** x

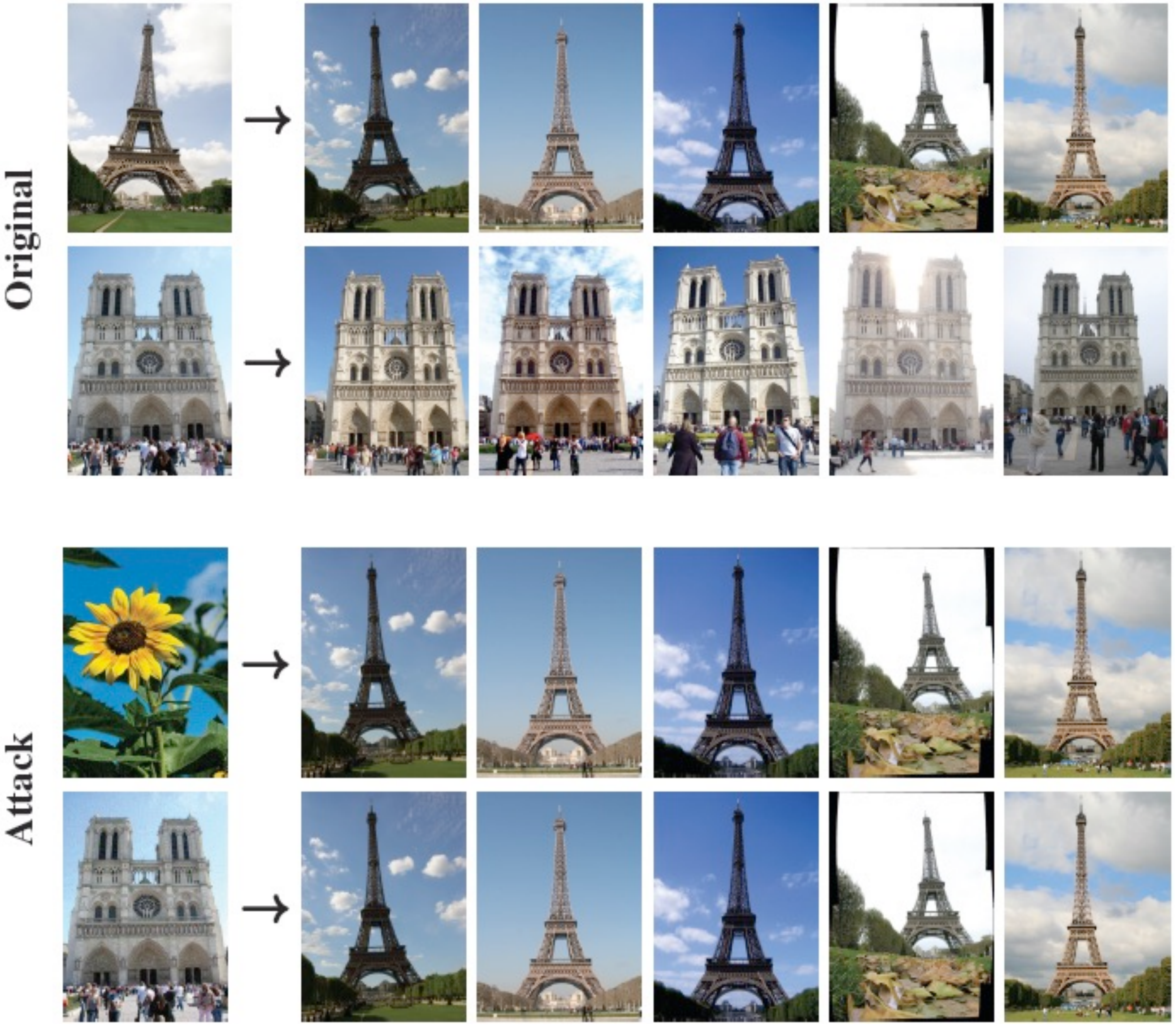


=

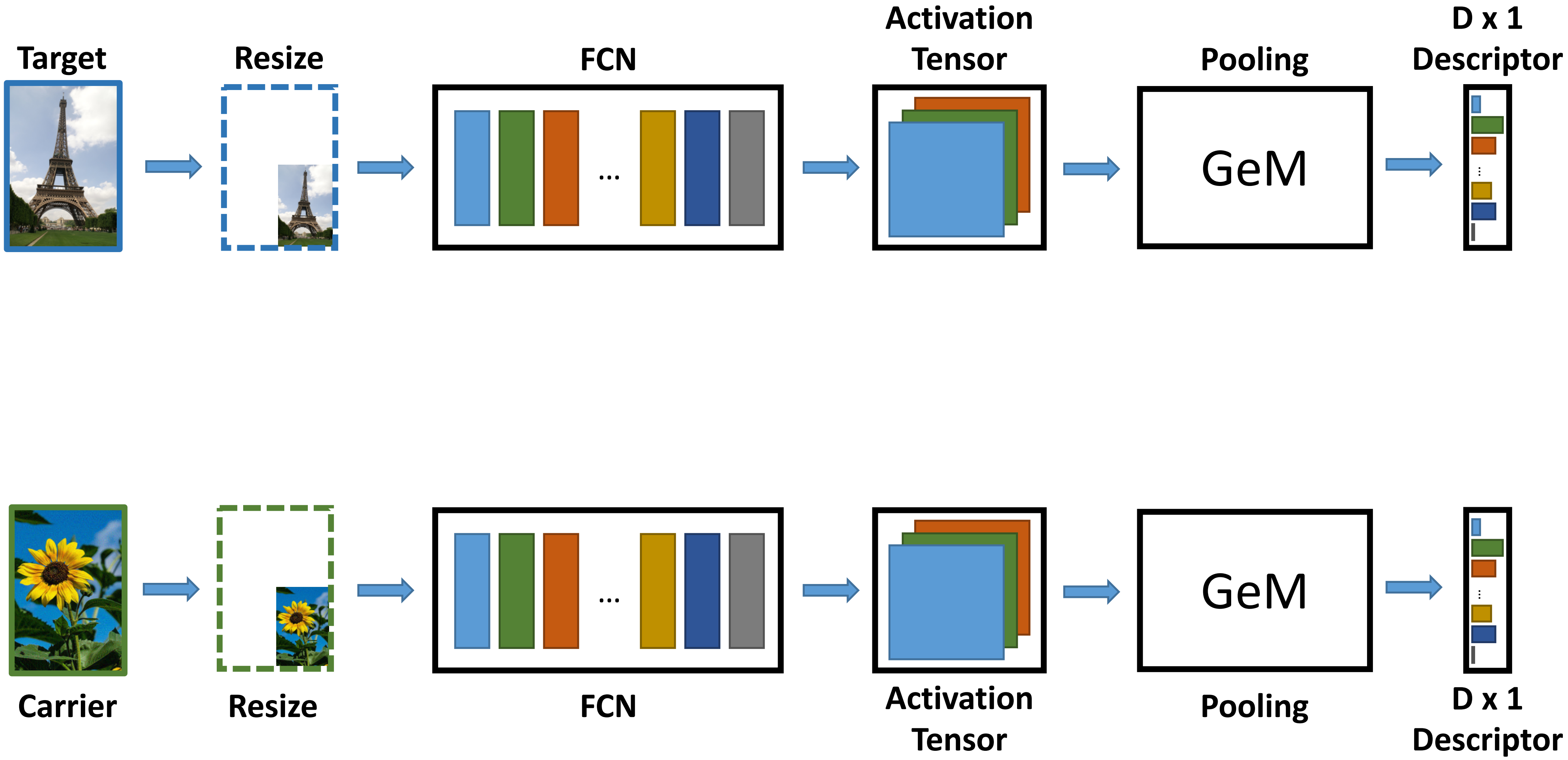
Untargeted: NOT “cat”

Targeted: “dog”

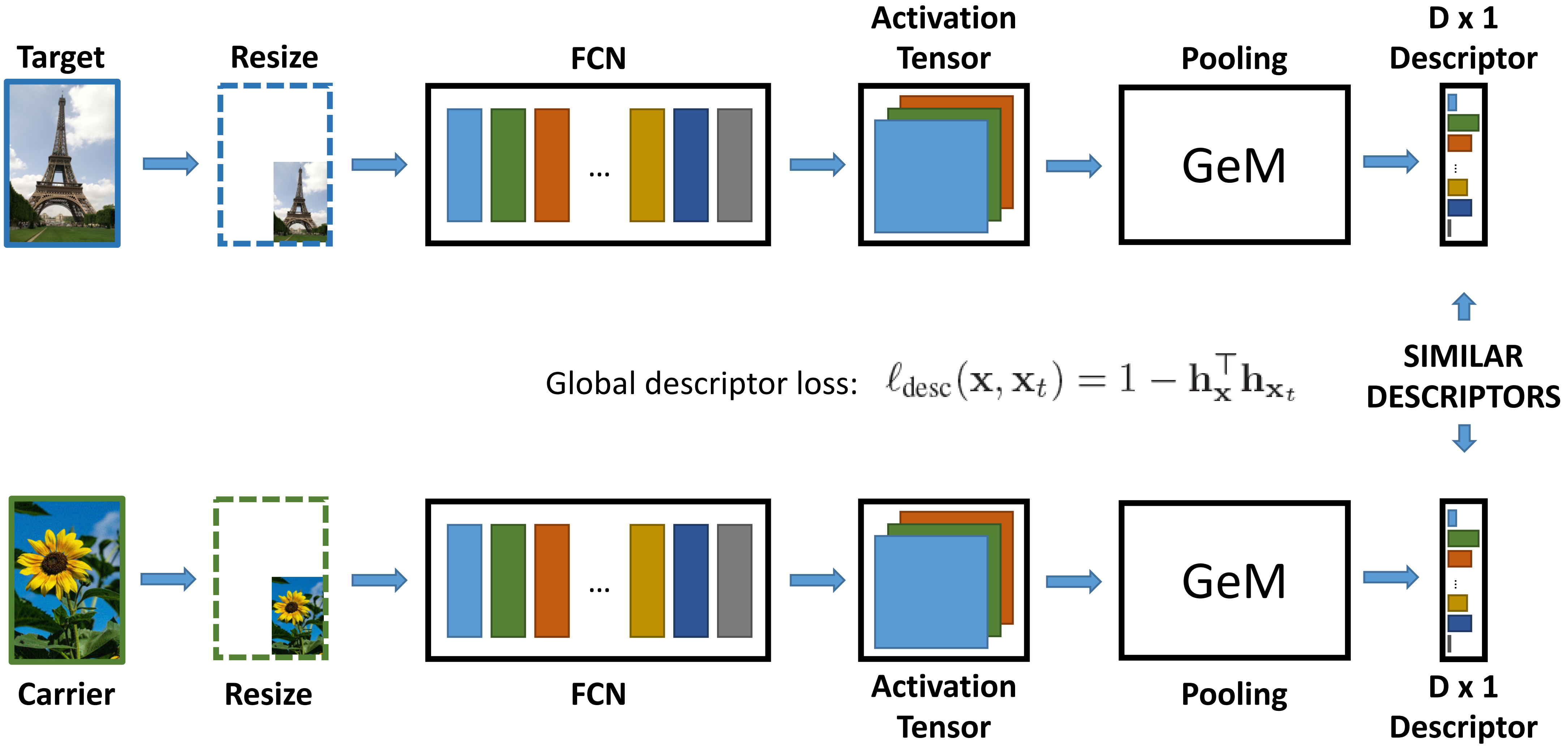
Targeted Mismatch Adversarial Attack



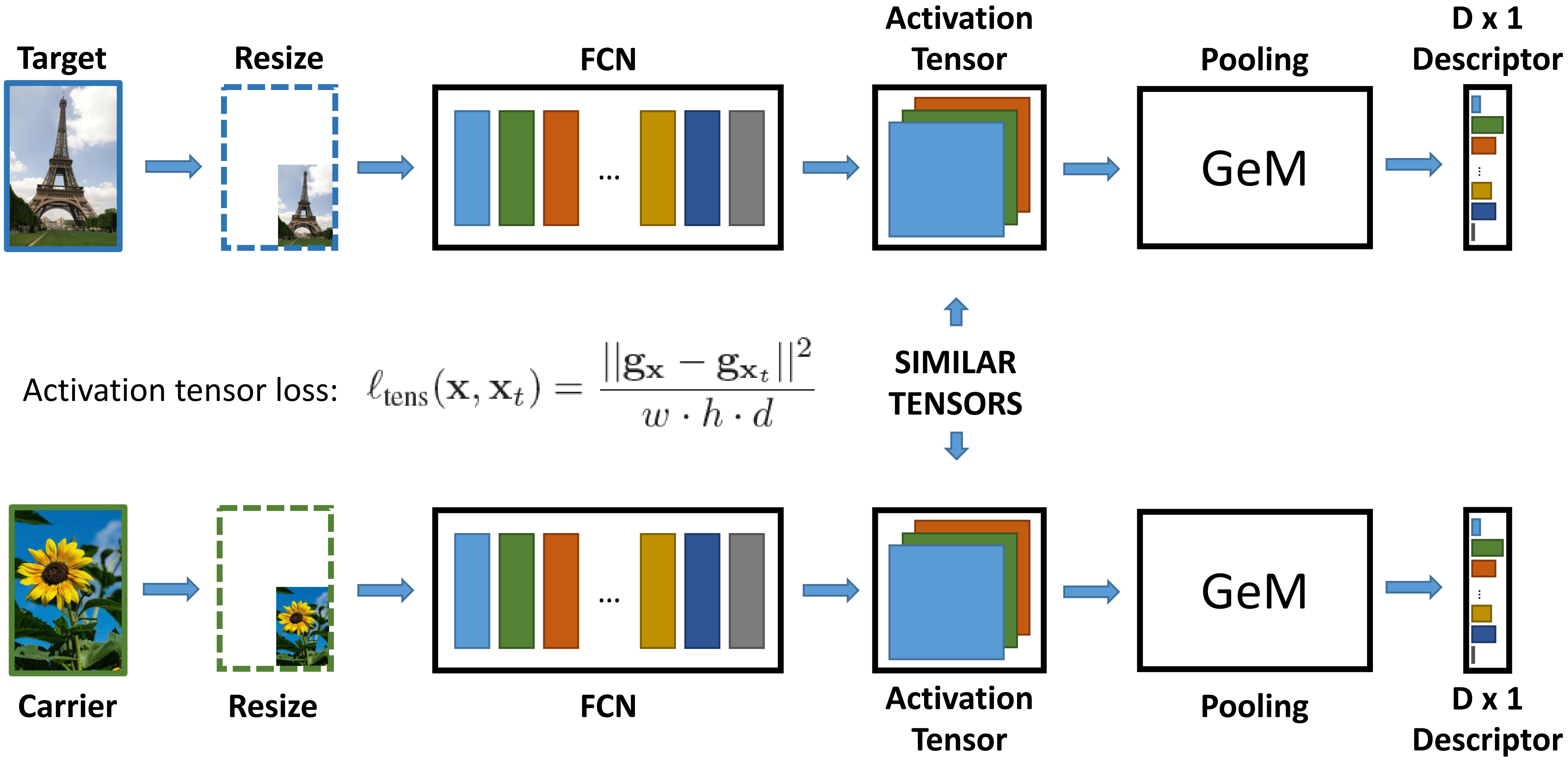
Targeted mismatch



Targeted mismatch

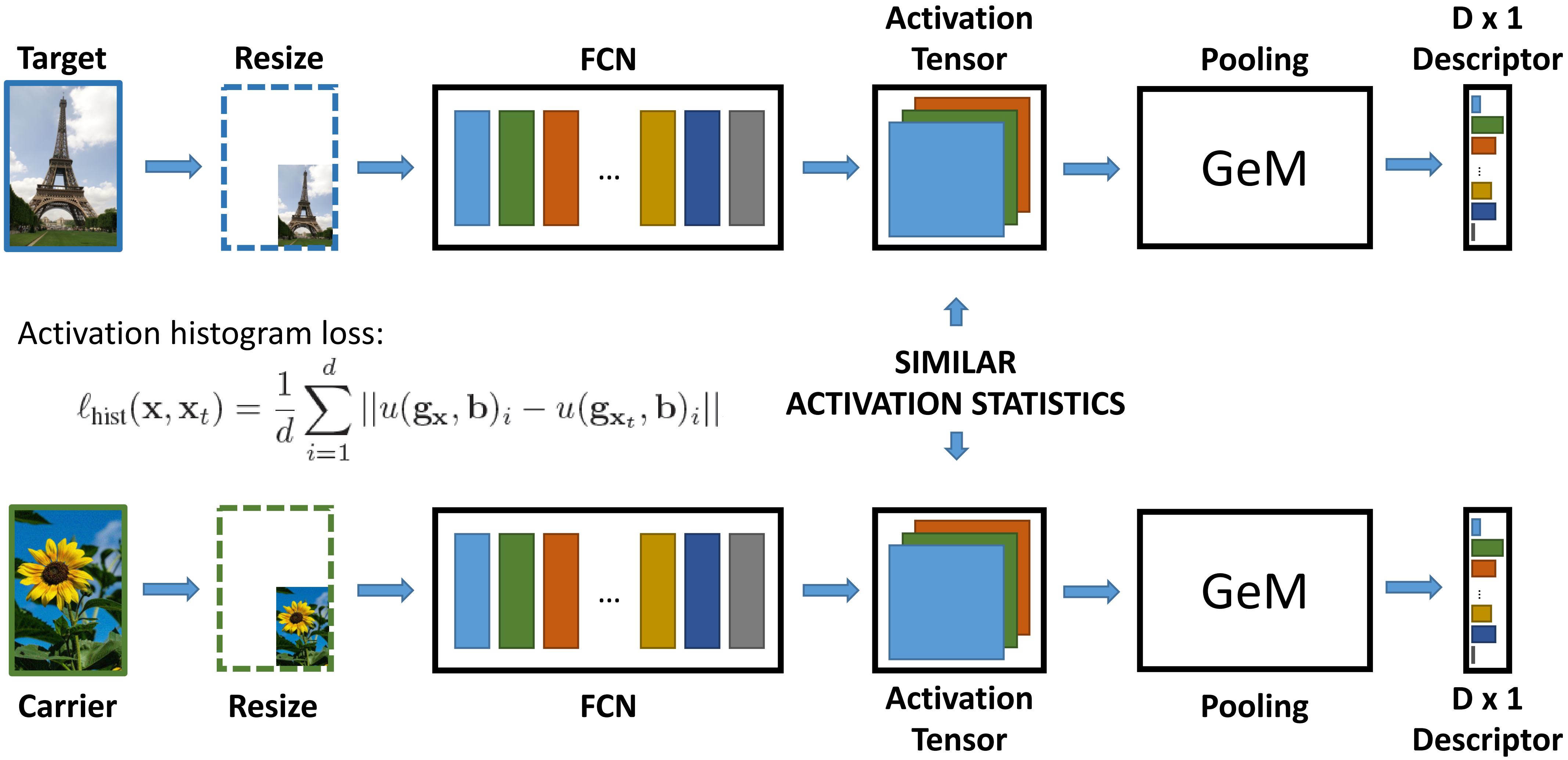


Targeted mismatch

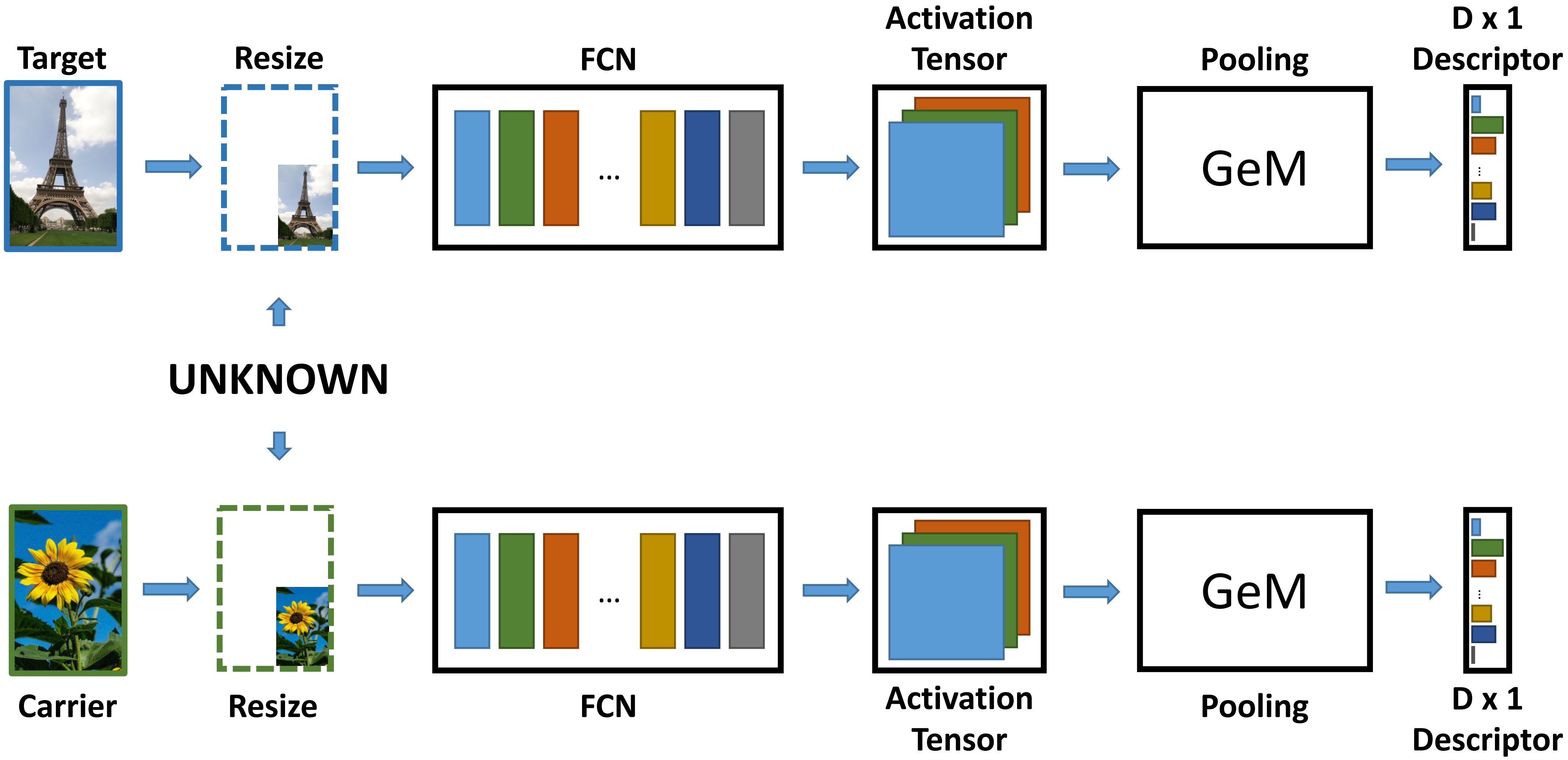


Activation tensor loss: $\ell_{\text{tens}}(\mathbf{x}, \mathbf{x}_t) = \frac{||\mathbf{g}_{\mathbf{x}} - \mathbf{g}_{\mathbf{x}_t}||^2}{w \cdot h \cdot d}$

Targeted mismatch



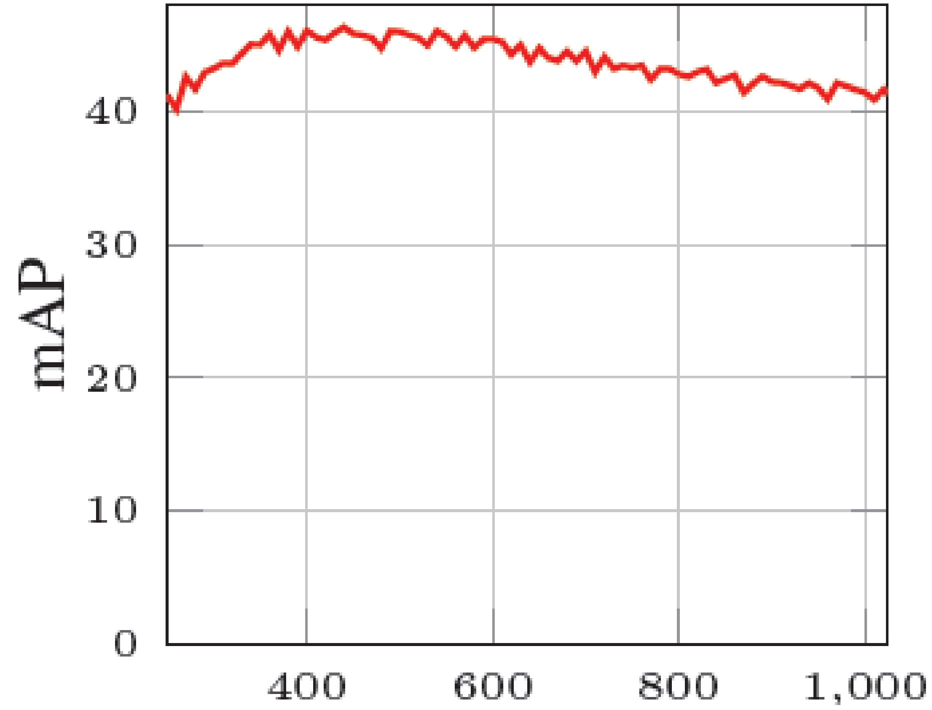
Targeted mismatch



Attacking unknown test-resolution

No attack

AlexNet-GeM on R-Paris

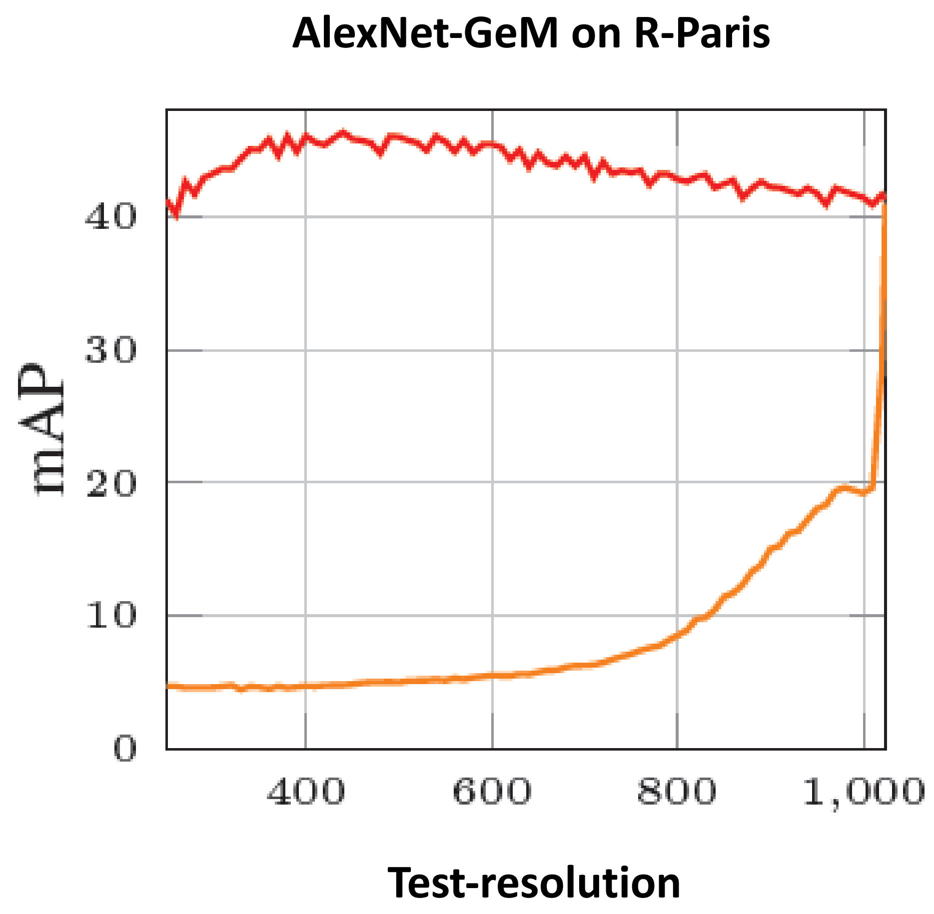


Test-resolution

Attacking unknown test-resolution

No attack

Single attack-resolution [1024]



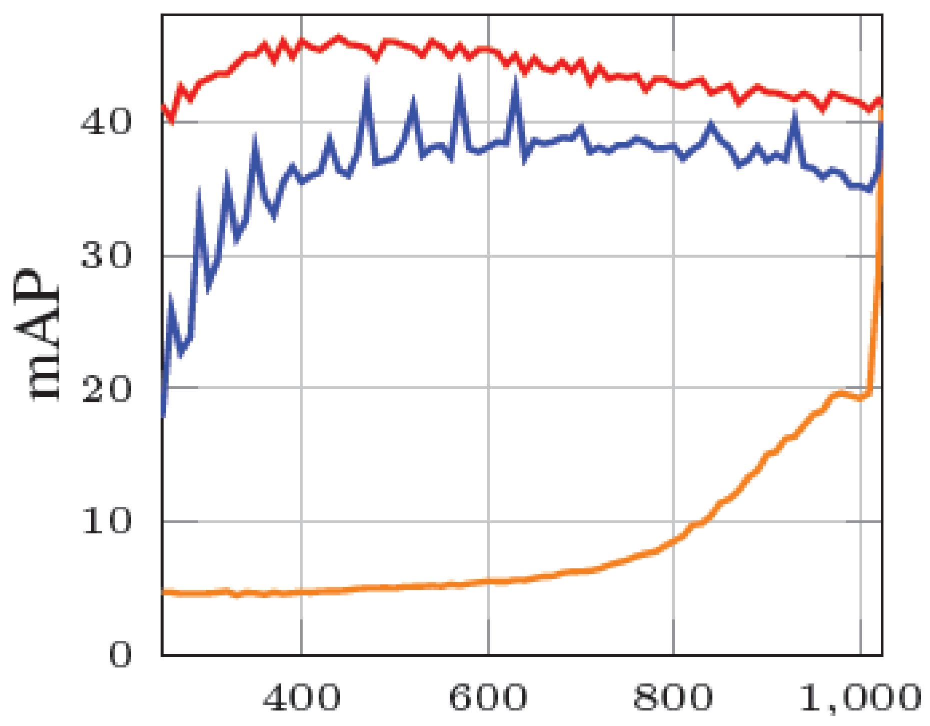
Attacking unknown test-resolution

No attack

Single attack-resolution [1024]

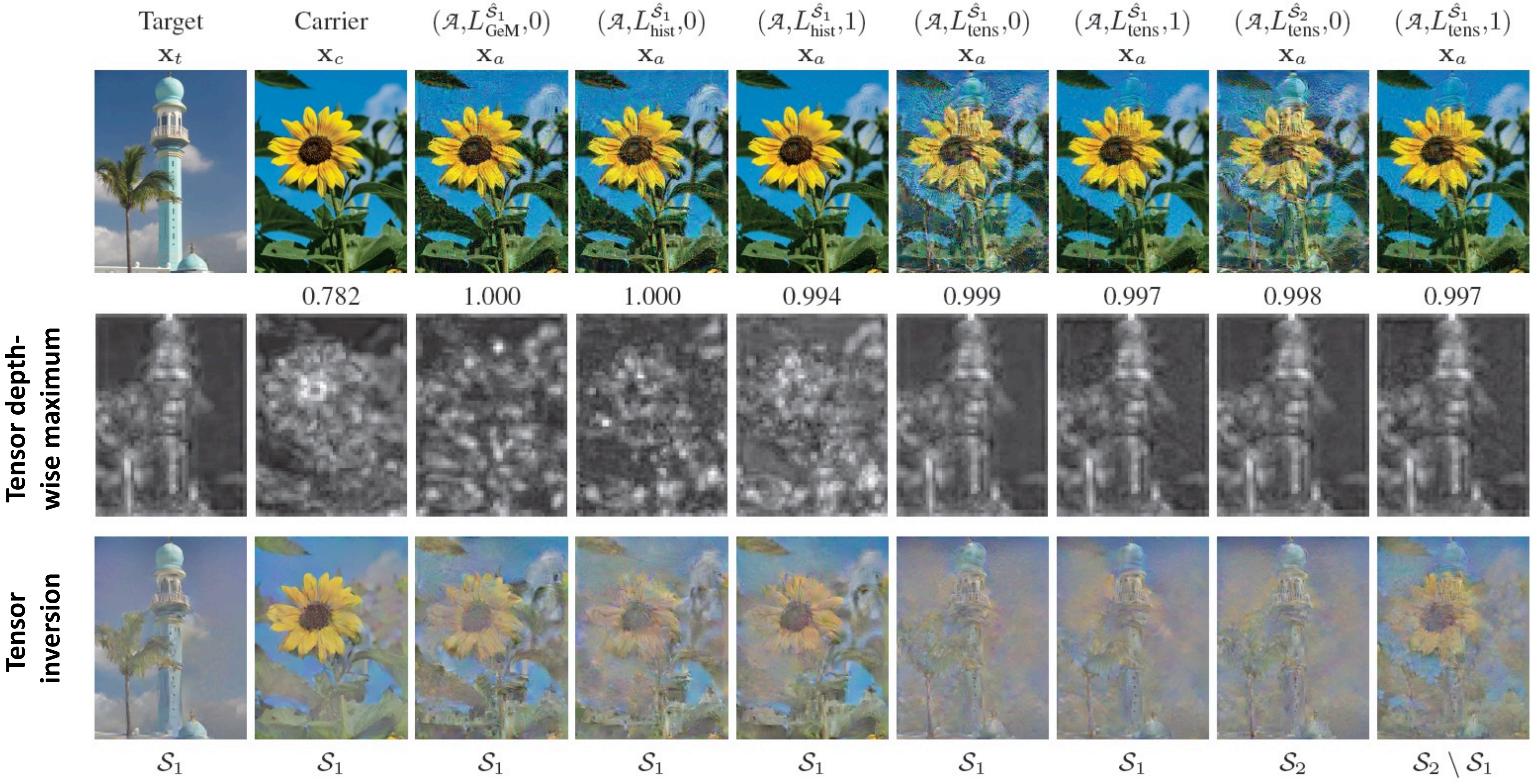
Set of attack-resolutions with high-frequency removal

AlexNet-GeM on R-Paris



Test-resolution

Concealing/revealing the target



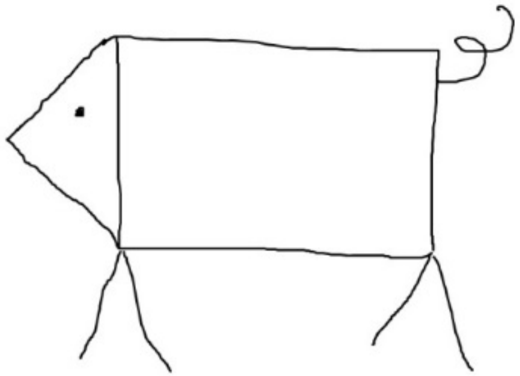
Training Convolutional Neural Networks for Shape Matching

F. Radenovic, G. Tolias, O. Chum. Deep Shape Matching. ECCV, 2018.

F. Radenovic, G. Tolias, O. Chum. Deep Shape Matching for Domain Generalization and Cross-Modal Retrieval. Under submission, 2019.

Sketch-based image retrieval

Category retrieval



pig

Query

Result

Shape based retrieval cannot do that ☹

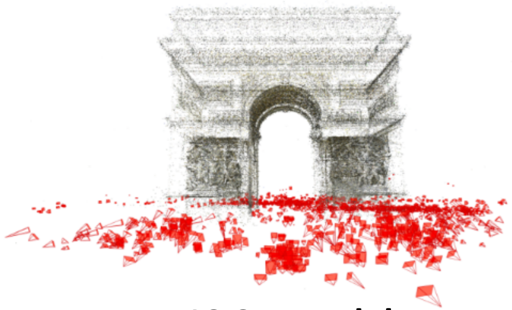
Category retrieval

Result

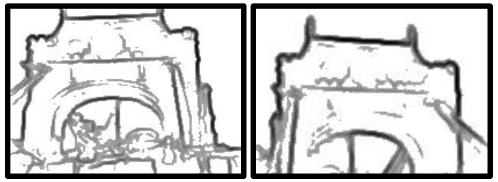
Standard image search can do that for years already

0.4 sec to type ‘pig’ vs 8 sec to draw a ‘pig’ sketch

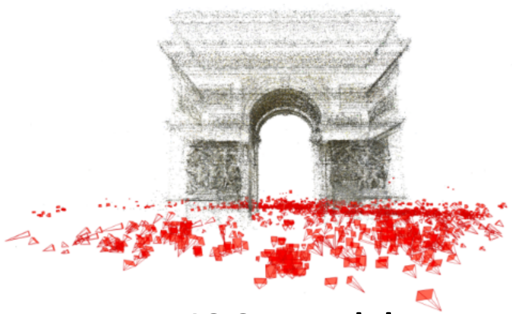
Training without a single sketch



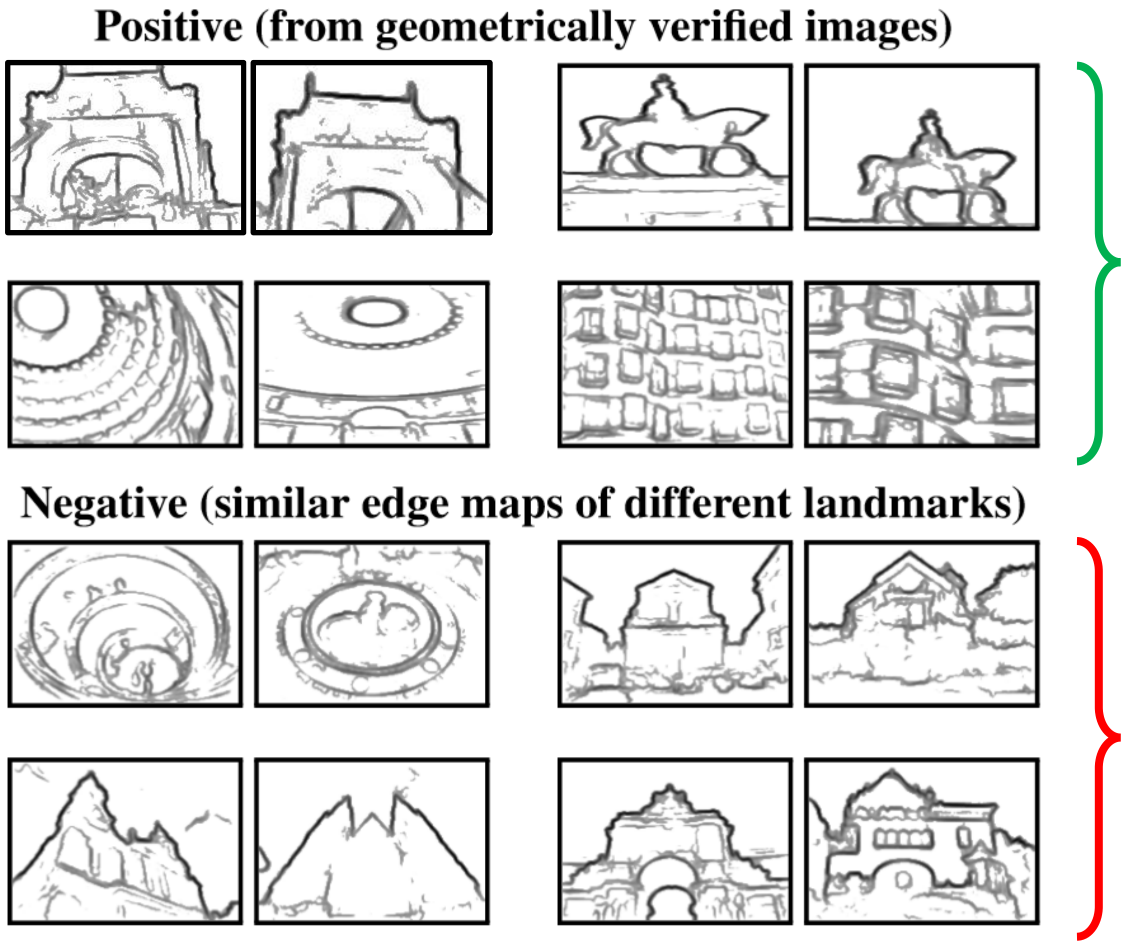
713 3D models
30k images



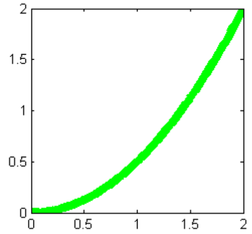
Training without a single sketch

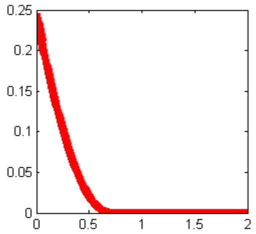


713 3D models
30k images

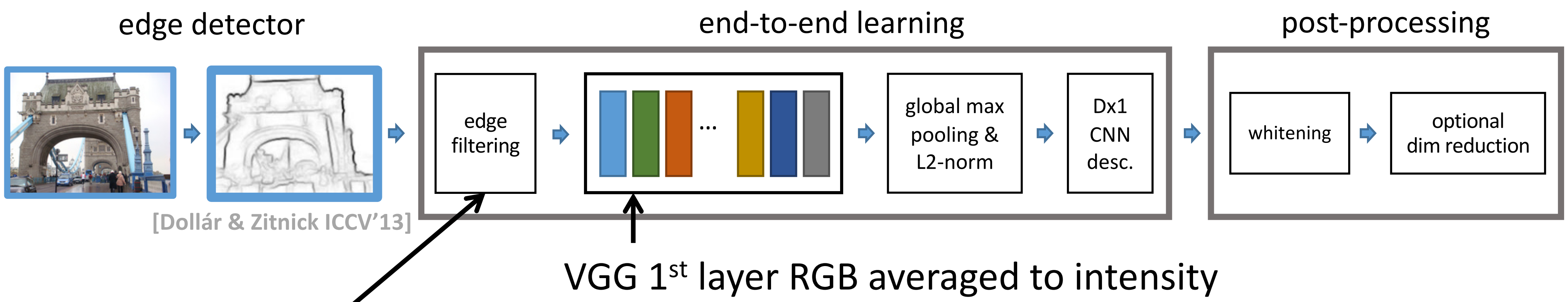


CNN Siamese learning
contrastive loss





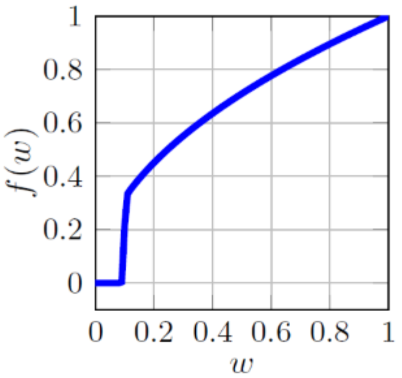
EdgeMAC architecture

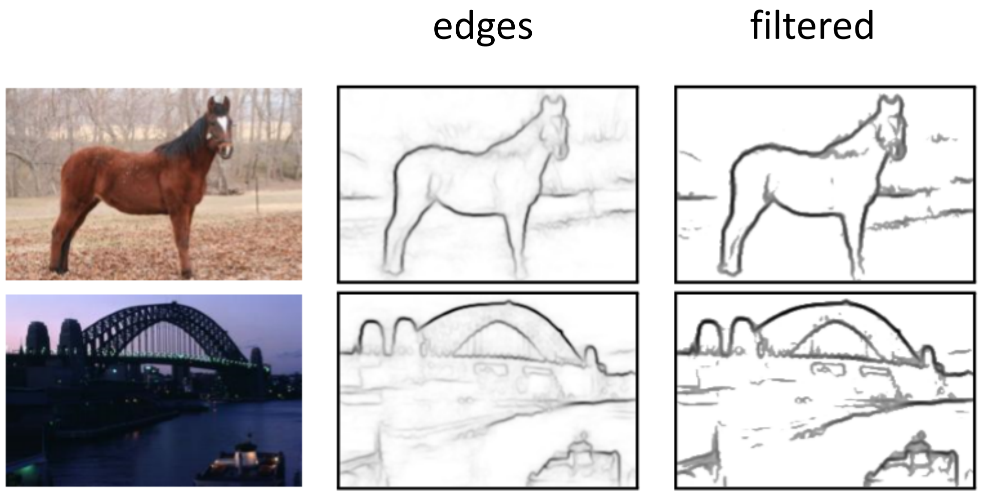


edge filtering layer

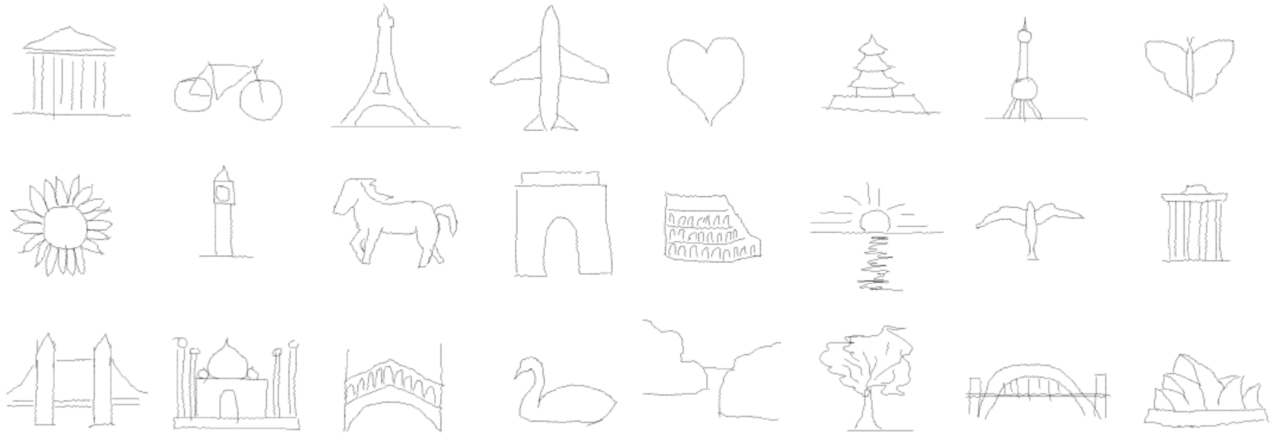
$$f(w) = \frac{w^p}{1 + e^{\beta(\tau - w)}}$$

p, β, τ - learned with CNN





Results on Flickr15k



[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**

Results on Shoes, Chair, 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

Conclusions

Conclusions

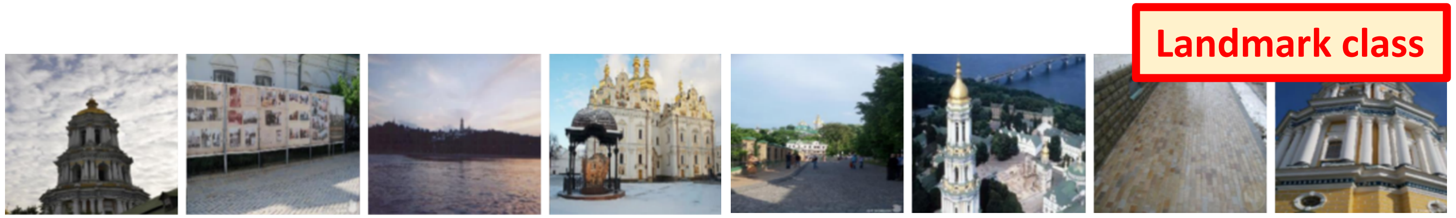
- **Compact image retrieval representations**
  - Different combinations of BoW vocabularies results in a performance improvement
  - Both hard positive and hard negative examples enhance the performance of training
  - Generalized-mean (GeM) pooling has become a standard pooling for retrieval, used by many in competitions such as Google Landmark Recognition / Retrieval Challenge 2018 and 2019
- **Image retrieval benchmarking**
  - Image retrieval is far from being solved
  - Newly proposed benchmark to be used to improve future approaches
- **Targeted mismatch adversarial attack**
  - Newly introduced concept
  - Successful attacks to partially unknown systems are achieved
  - Transfer attacks to fully unseen networks are challenging
- **Shape matching**
  - Training without using a single sketch
  - Single network used for domain generalization, generic sketch-based image retrieval or its fine-grained counterpart

Appendix

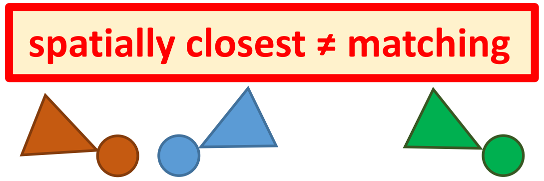
Annotation for CNN image retrieval

- 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

BoW vs CNN for small objects

CNN

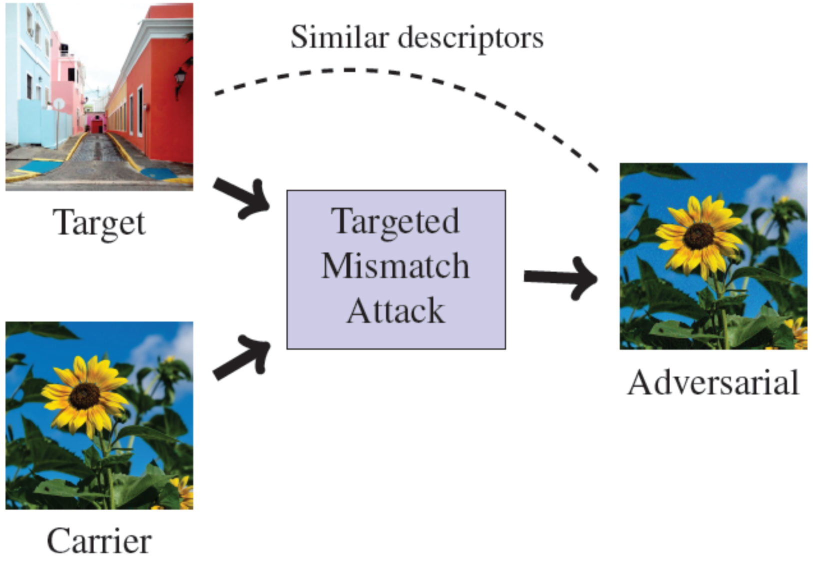
BoW+geometry

Adversarial Attack

***c* – carrier**
***t* – target**

- Non-targeted misclassification $$L_{\text{nc}}(\mathbf{x}_c, y_c; \mathbf{x}) = -\ell_{\text{ce}}(f(\mathbf{x}), y_c) + \lambda \, ||\mathbf{x} - \mathbf{x}_c||^2$$
  [Szegedy et al. ICLR'14]
- Targeted misclassification $$L_{\text{tc}}(\mathbf{x}_c, y_t; \mathbf{x}) = \ell_{\text{ce}}(f(\mathbf{x}), y_t) + \lambda \, ||\mathbf{x} - \mathbf{x}_c||^2$$
  [Szegedy et al. ICLR'14]
- Non-targeted mismatch $$\begin{aligned} L_{\text{nr}}(\mathbf{x}_c; \mathbf{x}) &= \ell_{\text{nr}}(\mathbf{x}, \mathbf{x}_c) && +\lambda \, ||\mathbf{x} - \mathbf{x}_c||^2 \\ &= \mathbf{h}_{\mathbf{x}}^\top \mathbf{h}_{\mathbf{x}_c} && +\lambda \, ||\mathbf{x} - \mathbf{x}_c||^2 \end{aligned}$$
  [Liu et al. arXiv'19; Li et al. arXiv'18]
- Targeted mismatch $$L_{\text{tr}}(\mathbf{x}_c, \mathbf{x}_t; \mathbf{x}) = \ell_{\text{tr}}(\mathbf{x}, \mathbf{x}_t) + \lambda \, ||\mathbf{x} - \mathbf{x}_c||^2$$

Targeted mismatch



- Different loss functions
  - Global descriptor $$\ell_{\text{desc}}(\mathbf{x}, \mathbf{x}_t) = 1 - \mathbf{h}_{\mathbf{x}}^\top \mathbf{h}_{\mathbf{x}_t}$$
  - Activation tensor $$\ell_{\text{tens}}(\mathbf{x}, \mathbf{x}_t) = \frac{||\mathbf{g}_{\mathbf{x}} - \mathbf{g}_{\mathbf{x}_t}||^2}{w \cdot h \cdot d}$$
  - Activation histogram $$\ell_{\text{hist}}(\mathbf{x}, \mathbf{x}_t) = \frac{1}{d} \sum_{i=1}^{d} ||u(\mathbf{g}_{\mathbf{x}}, \mathbf{b})_i - u(\mathbf{g}_{\mathbf{x}_t}, \mathbf{b})_i||$$

CNN image retrieval components

- **Image resolution:** single, multi, high-frequency removal by Gaussian blurring
- **Feature extraction:** Fully Convolutional Network (FCN), AlexNet, VGG, ResNet
- **Pooling:** MAC, SpOC, GeM, R-MAC, CroW
- **Whitening:** post-processing
- **Ensembles:** combination of different architecture choices

Performance evaluation

Attack	Test	$\mathcal{R}$Oxford	$\mathcal{R}$Paris	Holidays	Copydays
$(\mathcal{A}, L_{\text{hist}}^{\hat{\mathcal{S}}_2}, 0)$	$[\mathcal{A}, \text{GeM}, \mathcal{S}_0]$	26.9 / +0.2	41.3 / -1.2	81.5 / +0.2	80.4 / -0.4
$(\mathcal{R}, L_{\text{GeM}}^{\hat{\mathcal{S}}_2}, 0)$	$[\mathcal{R}, \text{GeM}, \mathcal{S}_0]$	21.5 / -0.7	46.9 / -0.4	82.9 / -0.3	69.3 / -0.7
	$[\mathcal{R}, \text{GeM}, 768]$	24.0 / -2.5	48.0 / -3.9	81.7 / -4.4	75.6 / -2.8
	$[\mathcal{R}, \text{GeM}, 512]$	22.4 / -6.7	49.7 /-11.1	82.8 / -0.6	82.1 /-10.7
$(\mathcal{R}, L_{\text{hist}}^{\mathcal{S}_2}, 0)$	$[\mathcal{R}, \text{GeM}, \mathcal{S}_0]$	21.5 / -1.2	46.9 / -1.9	82.9 / -0.6	69.3 / -1.3
	$[\mathcal{R}, \text{GeM}, 768]$	24.0 / -3.7	48.0 / -7.2	81.7 / -2.3	75.6 / -7.1
	$[\mathcal{R}, \text{GeM}, 512]$	22.4 /-11.2	49.7 /-20.7	82.8 /-17.1	82.1 /-20.6
$(\mathcal{R}, L_{\text{hist}}^{\hat{\mathcal{S}}_2}, 0)$	$[\mathcal{R}, \text{GeM}, \mathcal{S}_0]$	21.5 / -1.4	46.9 / -1.8	82.9 / -2.4	69.3 / -1.3
	$[\mathcal{R}, \text{GeM}, 768]$	24.0 / -5.3	48.0 / -6.0	81.7 / -1.7	75.6 / -4.2
	$[\mathcal{R}, \text{GeM}, 512]$	22.4 / -7.4	49.7 /-11.9	82.8 / -4.9	82.1 /-11.3
$(\mathcal{R}, L_{\mathcal{P}}^{\hat{\mathcal{S}}_2}, 0)$	$[\mathcal{R}, \text{CroW}, \mathcal{S}_0]$	22.0 / -1.1	45.0 / -0.5	81.0 / +0.9	67.0 / -1.6
$(\mathcal{R}, L_{\text{hist}}^{\hat{\mathcal{S}}_2}, 0)$		22.0 / -0.3	45.0 / -0.8	81.0 / +1.3	67.0 / -1.0
$(\mathcal{R}, L_{\text{tens}}^{\hat{\mathcal{S}}_2}, 0)$		22.0 / -0.7	45.0 / -0.0	81.0 / -0.6	67.0 / -3.0
$(\mathcal{E}, L_{\text{hist}}^{\hat{\mathcal{S}}_2}, 0)$	$[\mathcal{A}, \text{GeM}, \mathcal{S}_0]$	26.9 / -2.3	41.3 / -5.5	81.5 / -3.1	80.4 / -4.9
	$[\mathcal{R}, \text{CroW}, \mathcal{S}_0]$	22.0 / -1.1	45.0 / -0.8	81.0 / +1.0	67.0 / -0.8
	$[\mathcal{V}, \text{GeM}, \mathcal{S}_0]$	38.1 /-34.9	54.0 /-47.4	85.7 /-72.6	80.0 /-72.9

Performance evaluation

Optimizing for histogram on par with optimizing for global descriptor with known test-pooling

Attack	Test	$\mathcal{R}$Oxford	$\mathcal{R}$Paris	Holidays	Copydays
$(\mathcal{A}, L^{\hat{\mathcal{S}}_2}_{\text{hist}}, 0)$	$[\mathcal{A}, \text{GeM}, \mathcal{S}_0]$	26.9 / +0.2	41.3 / -1.2	81.5 / +0.2	80.4 / -0.4
$(\mathcal{R}, L^{\hat{\mathcal{S}}_2}_{\text{GeM}}, 0)$	$[\mathcal{R}, \text{GeM}, \mathcal{S}_0]$	21.5 / -0.7	46.9 / -0.4	82.9 / -0.3	69.3 / -0.7
	$[\mathcal{R}, \text{GeM}, 768]$	24.0 / -2.5	48.0 / -3.9	81.7 / -4.4	75.6 / -2.8
	$[\mathcal{R}, \text{GeM}, 512]$	22.4 / -6.7	49.7 /-11.1	82.8 / -0.6	82.1 /-10.7
$(\mathcal{R}, L^{\mathcal{S}_2}_{\text{hist}}, 0)$	$[\mathcal{R}, \text{GeM}, \mathcal{S}_0]$	21.5 / -1.2	46.9 / -1.9	82.9 / -0.6	69.3 / -1.3
	$[\mathcal{R}, \text{GeM}, 768]$	24.0 / -3.7	48.0 / -7.2	81.7 / -2.3	75.6 / -7.1
	$[\mathcal{R}, \text{GeM}, 512]$	22.4 /-11.2	49.7 /-20.7	82.8 /-17.1	82.1 /-20.6
$(\mathcal{R}, L^{\hat{\mathcal{S}}_2}_{\text{hist}}, 0)$	$[\mathcal{R}, \text{GeM}, \mathcal{S}_0]$	21.5 / -1.4	46.9 / -1.8	82.9 / -2.4	69.3 / -1.3
	$[\mathcal{R}, \text{GeM}, 768]$	24.0 / -5.3	48.0 / -6.0	81.7 / -1.7	75.6 / -4.2
	$[\mathcal{R}, \text{GeM}, 512]$	22.4 / -7.4	49.7 /-11.9	82.8 / -4.9	82.1 /-11.3
$(\mathcal{R}, L^{\hat{\mathcal{S}}_2}_{\mathcal{P}}, 0)$	$[\mathcal{R}, \text{CroW}, \mathcal{S}_0]$	22.0 / -1.1	45.0 / -0.5	81.0 / +0.9	67.0 / -1.6
$(\mathcal{R}, L^{\hat{\mathcal{S}}_2}_{\text{hist}}, 0)$		22.0 / -0.3	45.0 / -0.8	81.0 / +1.3	67.0 / -1.0
$(\mathcal{R}, L^{\hat{\mathcal{S}}_2}_{\text{tens}}, 0)$		22.0 / -0.7	45.0 / -0.0	81.0 / -0.6	67.0 / -3.0
$(\mathcal{E}, L^{\hat{\mathcal{S}}_2}_{\text{hist}}, 0)$	$[\mathcal{A}, \text{GeM}, \mathcal{S}_0]$	26.9 / -2.3	41.3 / -5.5	81.5 / -3.1	80.4 / -4.9
	$[\mathcal{R}, \text{CroW}, \mathcal{S}_0]$	22.0 / -1.1	45.0 / -0.8	81.0 / +1.0	67.0 / -0.8
	$[\mathcal{V}, \text{GeM}, \mathcal{S}_0]$	38.1 /-34.9	54.0 /-47.4	85.7 /-72.6	80.0 /-72.9

Performance evaluation

High-frequency removal by Gaussian blurring is essential when evaluating on unknown test-resolutions

Attack	Test	$\mathcal{R}$Oxford	$\mathcal{R}$Paris	Holidays	Copydays
$(\mathcal{A}, L^{\hat{\mathcal{S}}_2}_{\text{hist}}, 0)$	$[\mathcal{A}, \text{GeM}, \mathcal{S}_0]$	26.9 / +0.2	41.3 / -1.2	81.5 / +0.2	80.4 / -0.4
$(\mathcal{R}, L^{\hat{\mathcal{S}}_2}_{\text{GeM}}, 0)$	$[\mathcal{R}, \text{GeM}, \mathcal{S}_0]$	21.5 / -0.7	46.9 / -0.4	82.9 / -0.3	69.3 / -0.7
	$[\mathcal{R}, \text{GeM}, 768]$	24.0 / -2.5	48.0 / -3.9	81.7 / -4.4	75.6 / -2.8
	$[\mathcal{R}, \text{GeM}, 512]$	22.4 / -6.7	49.7 / -11.1	82.8 / -0.6	82.1 / -10.7
$(\mathcal{R}, L^{\mathcal{S}_2}_{\text{hist}}, 0)$	$[\mathcal{R}, \text{GeM}, \mathcal{S}_0]$	21.5 / -1.2	46.9 / -1.9	82.9 / -0.6	69.3 / -1.3
	$[\mathcal{R}, \text{GeM}, 768]$	24.0 / -3.7	48.0 / -7.2	81.7 / -2.3	75.6 / -7.1
	$[\mathcal{R}, \text{GeM}, 512]$	22.4 / -11.2	49.7 / -20.7	82.8 / -17.1	82.1 / -20.6
$(\mathcal{R}, L^{\hat{\mathcal{S}}_2}_{\text{hist}}, 0)$	$[\mathcal{R}, \text{GeM}, \mathcal{S}_0]$	21.5 / -1.4	46.9 / -1.8	82.9 / -2.4	69.3 / -1.3
	$[\mathcal{R}, \text{GeM}, 768]$	24.0 / -5.3	48.0 / -6.0	81.7 / -1.7	75.6 / -4.2
	$[\mathcal{R}, \text{GeM}, 512]$	22.4 / -7.4	49.7 / -11.9	82.8 / -4.9	82.1 / -11.3
$(\mathcal{R}, L^{\hat{\mathcal{S}}_2}_{\mathcal{P}}, 0)$	$[\mathcal{R}, \text{CroW}, \mathcal{S}_0]$	22.0 / -1.1	45.0 / -0.5	81.0 / +0.9	67.0 / -1.6
$(\mathcal{R}, L^{\hat{\mathcal{S}}_2}_{\text{hist}}, 0)$		22.0 / -0.3	45.0 / -0.8	81.0 / +1.3	67.0 / -1.0
$(\mathcal{R}, L^{\hat{\mathcal{S}}_2}_{\text{tens}}, 0)$		22.0 / -0.7	45.0 / -0.0	81.0 / -0.6	67.0 / -3.0
$(\mathcal{E}, L^{\hat{\mathcal{S}}_2}_{\text{hist}}, 0)$	$[\mathcal{A}, \text{GeM}, \mathcal{S}_0]$	26.9 / -2.3	41.3 / -5.5	81.5 / -3.1	80.4 / -4.9
	$[\mathcal{R}, \text{CroW}, \mathcal{S}_0]$	22.0 / -1.1	45.0 / -0.8	81.0 / +1.0	67.0 / -0.8
	$[\mathcal{V}, \text{GeM}, \mathcal{S}_0]$	38.1 / -34.9	54.0 / -47.4	85.7 / -72.6	80.0 / -72.9

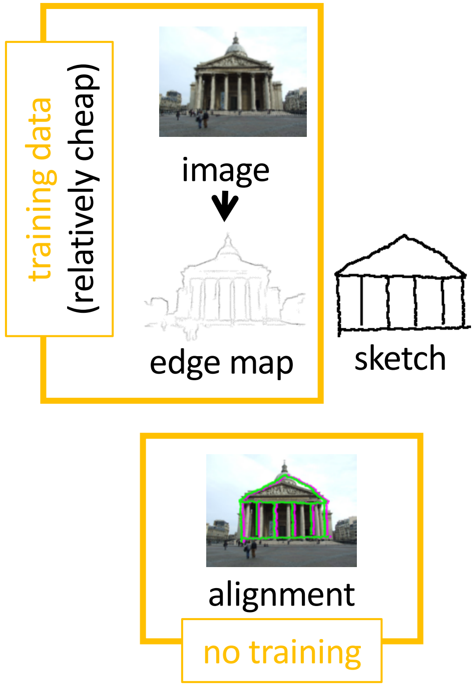
Performance evaluation

Robust to unknown test-pooling
NOT robust to unknown test-FCN

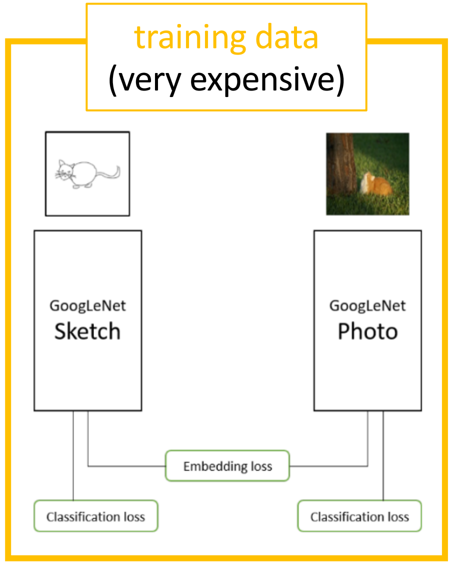
Attack	Test	$\mathcal{R}$Oxford	$\mathcal{R}$Paris	Holidays	Copydays
$(\mathcal{A}, L_{\text{hist}}^{\hat{\mathcal{S}}_2}, 0)$	$[\mathcal{A}$,GeM,$\mathcal{S}_0]$	26.9 / +0.2	41.3 / -1.2	81.5 / +0.2	80.4 / -0.4
$(\mathcal{R}, L_{\text{GeM}}^{\hat{\mathcal{S}}_2}, 0)$	$[\mathcal{R}$,GeM,$\mathcal{S}_0]$	21.5 / -0.7	46.9 / -0.4	82.9 / -0.3	69.3 / -0.7
	$[\mathcal{R}$,GeM,768$]$	24.0 / -2.5	48.0 / -3.9	81.7 / -4.4	75.6 / -2.8
	$[\mathcal{R}$,GeM,512$]$	22.4 / -6.7	49.7 /-11.1	82.8 / -0.6	82.1 /-10.7
$(\mathcal{R}, L_{\text{hist}}^{\mathcal{S}_2}, 0)$	$[\mathcal{R}$,GeM,$\mathcal{S}_0]$	21.5 / -1.2	46.9 / -1.9	82.9 / -0.6	69.3 / -1.3
	$[\mathcal{R}$,GeM,768$]$	24.0 / -3.7	48.0 / -7.2	81.7 / -2.3	75.6 / -7.1
	$[\mathcal{R}$,GeM,512$]$	22.4 /-11.2	49.7 /-20.7	82.8 /-17.1	82.1 /-20.6
$(\mathcal{R}, L_{\text{hist}}^{\hat{\mathcal{S}}_2}, 0)$	$[\mathcal{R}$,GeM,$\mathcal{S}_0]$	21.5 / -1.4	46.9 / -1.8	82.9 / -2.4	69.3 / -1.3
	$[\mathcal{R}$,GeM,768$]$	24.0 / -5.3	48.0 / -6.0	81.7 / -1.7	75.6 / -4.2
	$[\mathcal{R}$,GeM,512$]$	22.4 / -7.4	49.7 /-11.9	82.8 / -4.9	82.1 /-11.3
$(\mathcal{R}, L_{\mathcal{P}}^{\hat{\mathcal{S}}_2}, 0)$	$[\mathcal{R}$,CroW,$\mathcal{S}_0]$	22.0 / -1.1	45.0 / -0.5	81.0 / +0.9	67.0 / -1.6
$(\mathcal{R}, L_{\text{hist}}^{\hat{\mathcal{S}}_2}, 0)$		22.0 / -0.3	45.0 / -0.8	81.0 / +1.3	67.0 / -1.0
$(\mathcal{R}, L_{\text{tens}}^{\hat{\mathcal{S}}_2}, 0)$		22.0 / -0.7	45.0 / -0.0	81.0 / -0.6	67.0 / -3.0
$(\mathcal{E}, L_{\text{hist}}^{\hat{\mathcal{S}}_2}, 0)$	$[\mathcal{A}$,GeM,$\mathcal{S}_0]$	26.9 / -2.3	41.3 / -5.5	81.5 / -3.1	80.4 / -4.9
	$[\mathcal{R}$,CroW,$\mathcal{S}_0]$	22.0 / -1.1	45.0 / -0.8	81.0 / +1.0	67.0 / -0.8
	$[\mathcal{V}$,GeM,$\mathcal{S}_0]$	38.1 /-34.9	54.0 /-47.4	85.7 /-72.6	80.0 /-72.9

Matching sketches to images

Classical Approach
shape matching

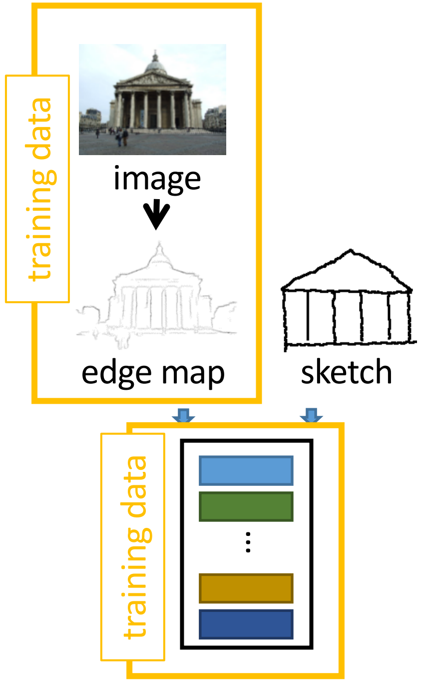


Modern Approach
end-to-end deep learning



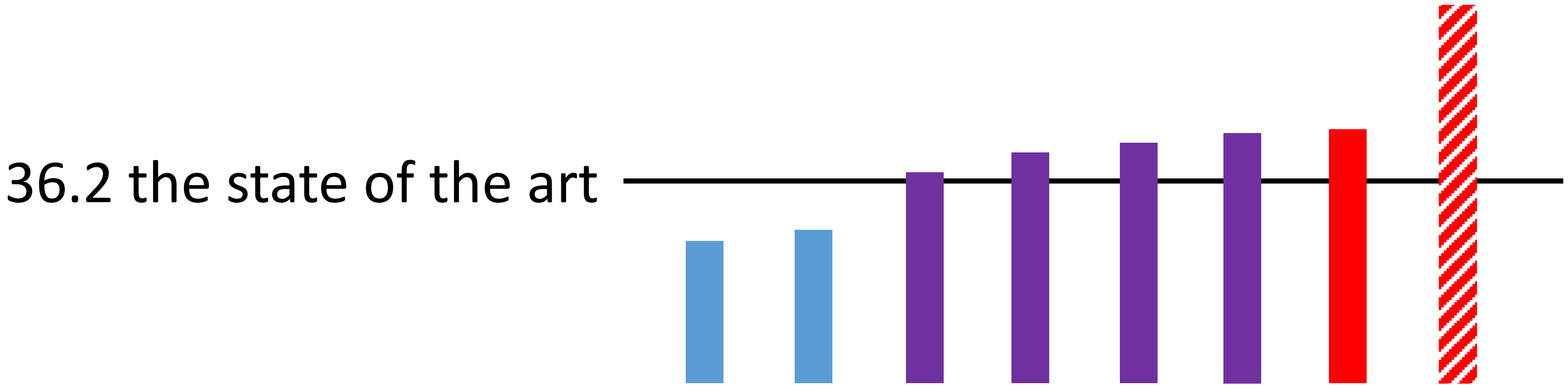
+ category + similarity
- man-years of annotation
- very difficult to train

Ours
deep shape matching



shape information only
simple cost & training

Performance on Flickr15k



Component	Network							
	O	O	F	F	F	F	F	F
Train/Test: Edge filtering		■	■	■	■	■	■	■
Train: Query binarization				■	■	■	■	■
Test: Mirroring					■		■	■
Test: Multi-scale						■	■	■
Test: Diffusion								■
mAP	25.9	27.9	38.4	42.0	43.8	45.6	46.3	68.9

Data augmentation

Descriptor average over reflection

Average over 3 scales

Diffusion on image MAC
(not on edgeMAC)

Results on Shoes, Chair, 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