

# Exploring and Learning from Visual Data

## Habilitation à Diriger des Recherches

Yannis Avrithis

Inria Rennes-Bretagne Atlantique

Rennes, July 2020

### Jury

Patrick Pérez - valeo.ai

Gabriela Csurka Khedari - Naver Labs

Jiri Matas - CTU Prague

Cordelia Schmid - Inria

Horst Bischof - TU Graz

Rémi Gribonval - Inria

Nikos Paragios - CentraleSupélec

Eric Marchand - UR1



# students and collaborators



Laurent Amsaleg



Mateusz Budnik



Andrei Bursuc



Ondrej Chum



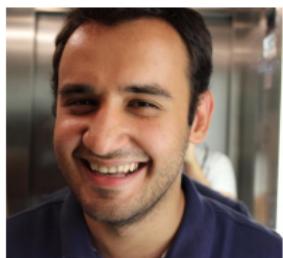
Ioannis Emiris



Teddy Furon

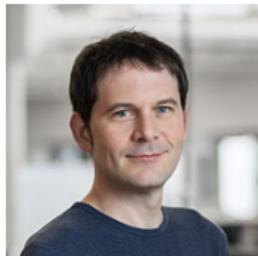


Guillaume Gravier



Ahmet Iscen

# students and collaborators



Hervé Jégou



Frédéric Jurie



Yannis Kalantidis



Ewa Kijak



Kimon Kontosis



Yann Lifchitz

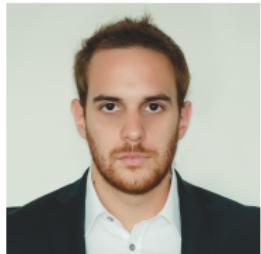


Marios Phinikettos



Sylvaine Picard

# students and collaborators



Filip Radenović



Kostas Rapantzikos



Miaojing Shi



Ronan Sicre



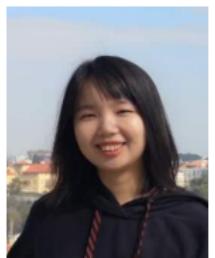
Oriane Siméoni



Giorgos Tolias

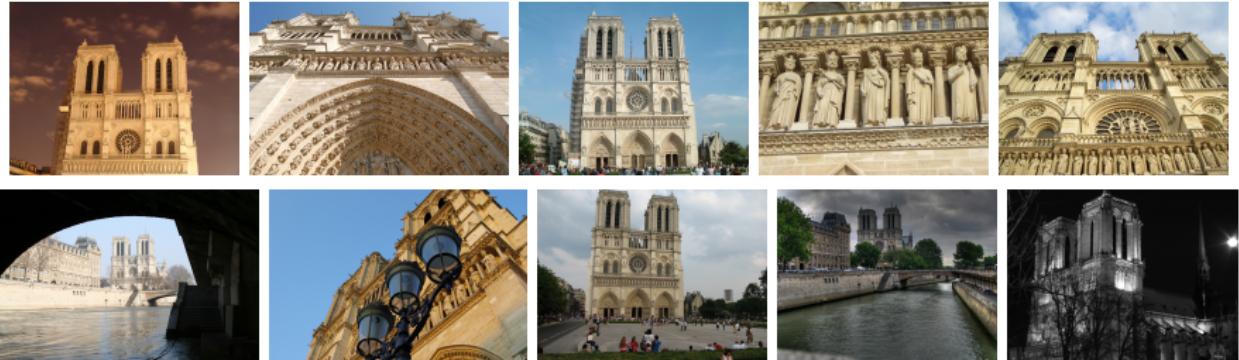


Christos Varytimidis

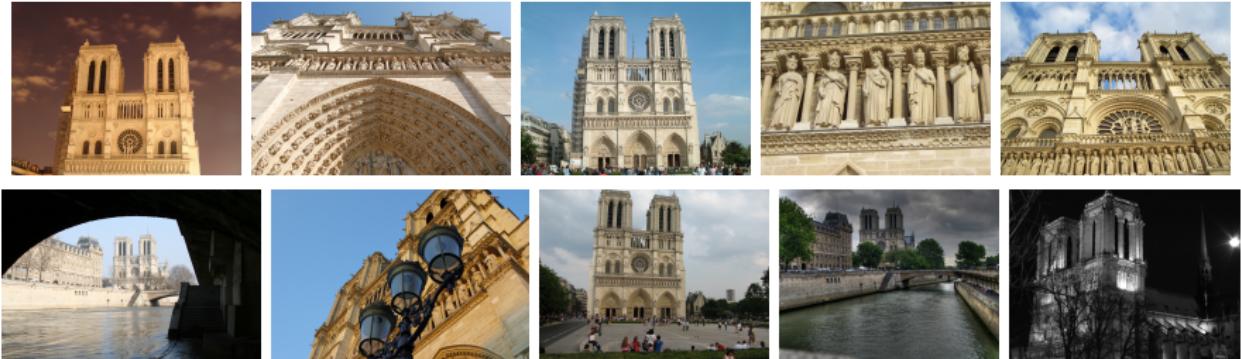


Hanwei Zhang

# instance-level tasks

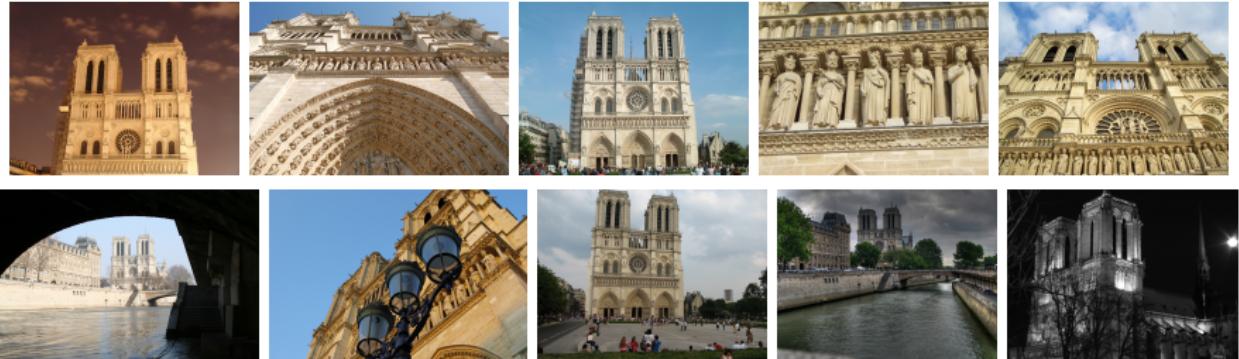


# instance-level tasks



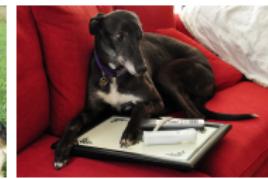
- scale
- viewpoint
- occlusion
- background clutter
- lighting

# instance-level tasks

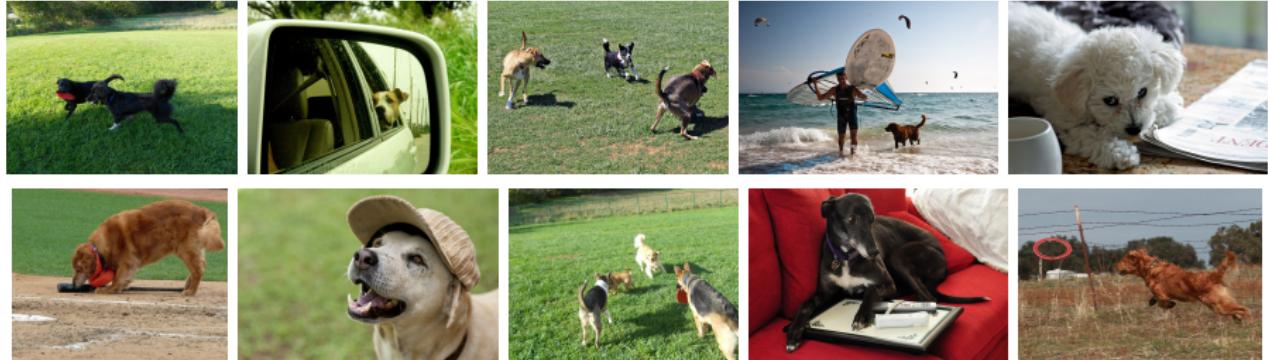


- scale
- viewpoint
- occlusion
- background clutter
- lighting
- discriminative power
- distractors

# category-level tasks



# category-level tasks



- scale
- viewpoint
- occlusion
- background clutter
- lighting

# category-level tasks



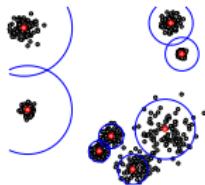
- scale
- viewpoint
- occlusion
- background clutter
- lighting
- number of instances
- texture/color
- pose
- deformability
- intra-class variability

## part I: exploring

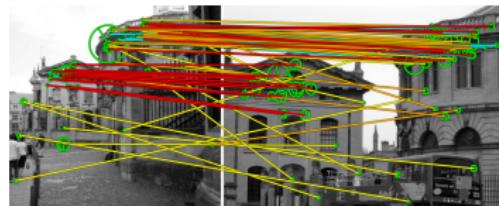
- instance-level visual matching, search and clustering
- shallow visual representations and matching processes
- local features, hand-crafted descriptors and visual vocabularies

## part I: exploring

- instance-level visual matching, search and clustering
- shallow visual representations and matching processes
- local features, hand-crafted descriptors and visual vocabularies



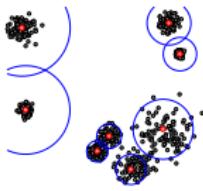
visual vocabularies



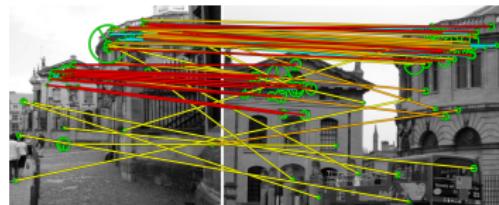
spatial matching

# part I: exploring

- instance-level visual matching, search and clustering
- shallow visual representations and matching processes
- local features, hand-crafted descriptors and visual vocabularies



visual vocabularies



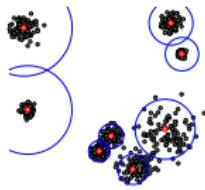
spatial matching



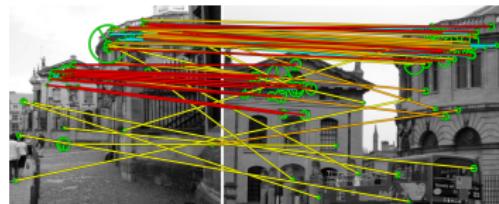
beyond vocabularies

# part I: exploring

- instance-level **visual matching**, **search** and **clustering**
- **shallow** visual representations and matching processes
- local features, hand-crafted descriptors and visual vocabularies



visual vocabularies



spatial matching



beyond vocabularies



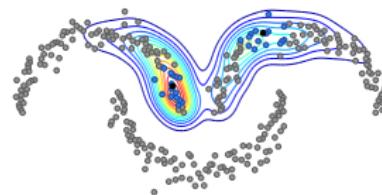
community photos

## part II: exploring deeper

- instance-level visual matching, search and object discovery
- deep visual representations and matching processes
- parametric models learned from visual data
- focus on the manifold structure of the feature space

## part II: exploring deeper

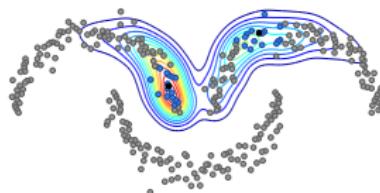
- instance-level visual matching, search and object discovery
- deep visual representations and matching processes
- parametric models learned from visual data
- focus on the manifold structure of the feature space



manifold search

## part II: exploring deeper

- instance-level visual matching, search and object discovery
- deep visual representations and matching processes
- parametric models learned from visual data
- focus on the manifold structure of the feature space



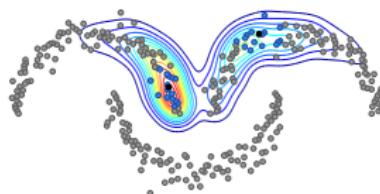
manifold search



spatial matching

## part II: exploring deeper

- instance-level visual matching, search and object discovery
- deep visual representations and matching processes
- parametric models learned from visual data
- focus on the manifold structure of the feature space



manifold search



spatial matching



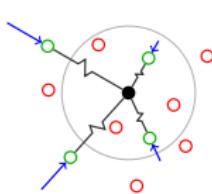
object discovery

## part III: learning

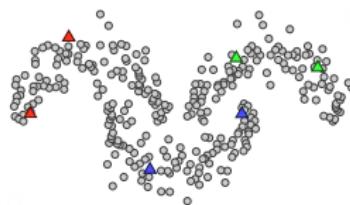
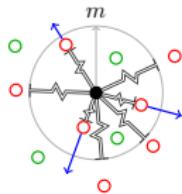
- learning deep visual representations by exploring visual data
- focus limited or no supervision
- progress from instance-level to category-level tasks

## part III: learning

- learning deep visual representations by exploring visual data
- focus limited or no supervision
- progress from instance-level to category-level tasks



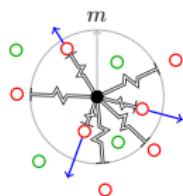
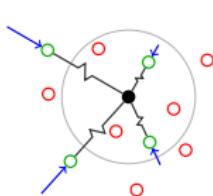
unsupervised metric learning



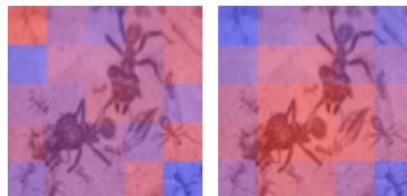
semi-supervised learning

## part III: learning

- learning deep visual representations by exploring visual data
- focus limited or no supervision
- progress from instance-level to category-level tasks



unsupervised metric learning      semi-supervised learning



few-shot learning

## part IV: beyond

### reflection

- current work
- take home message

### outlook

- a vision
- research directions

part I

# exploring

# outline – part I

2 context

3 visual vocabularies

4 spatial matching

5 beyond vocabularies

6 exploring photo collections

# scale-invariant feature transform (SIFT)



visual recognition works under occlusion, lighting and viewpoint changes

local feature  
detection by DoG

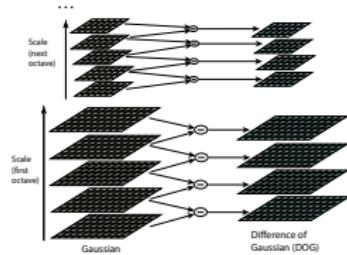
descriptor as histogram  
of gradient orientation

localization by  
Hough transform

# scale-invariant feature transform (SIFT)



visual recognition works under occlusion, lighting and viewpoint changes



local feature  
detection by DoG

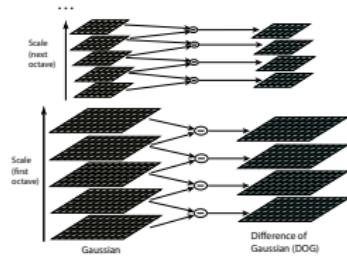
descriptor as histogram  
of gradient orientation

localization by  
Hough transform

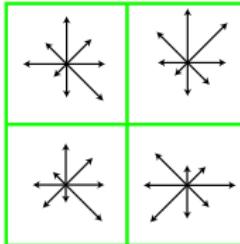
# scale-invariant feature transform (SIFT)



visual recognition works under occlusion, lighting and viewpoint changes



local feature detection by DoG



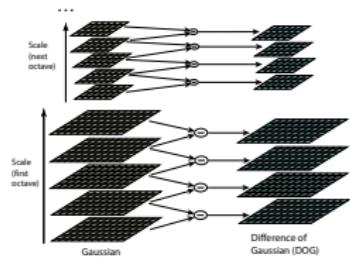
localization by Hough transform

Daugman. VR 1980. Two-Dimensional Spectral Analysis of Cortical Receptive Field Profiles.  
Lowe. ICCV 1999. Object recognition from local scale-invariant features.

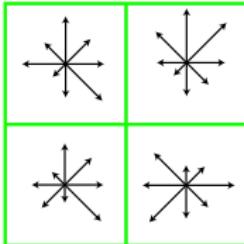
# scale-invariant feature transform (SIFT)



visual recognition works under occlusion, lighting and viewpoint changes



local feature  
detection by DoG



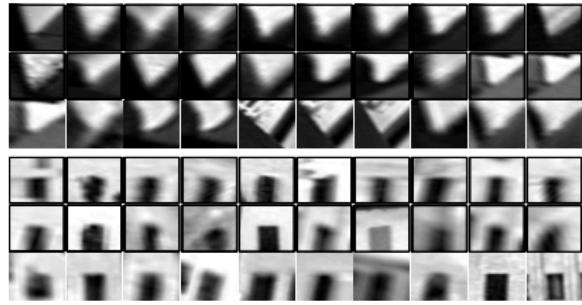
descriptor as histogram  
of gradient orientation



localization by  
Hough transform

Ballard. PR 1981. Generalizing the Hough Transform to Detect Arbitrary shapes.  
Lowe. ICCV 1999. Object recognition from local scale-invariant features.

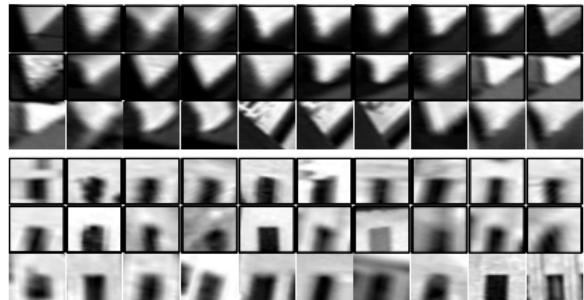
# bag of words (BoW)



## instance-level

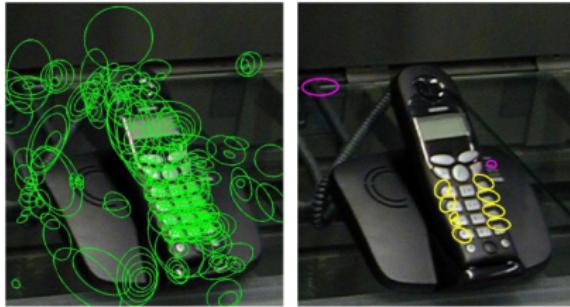
- clusters of SIFT descriptors
- images described by visual word histograms
- text retrieval, e.g. TF-IDF, inverted files

# bag of words (BoW)



## instance-level

- clusters of SIFT descriptors
- images described by visual word histograms
- text retrieval, e.g. TF-IDF, inverted files



## category-level

- naïve Bayes or SVM classifier
- features soon to be replaced by dense

Sivic and Zisserman. ICCV 2003. Video Google: A Text Retrieval Approach to Object Matching in videos.  
Csurka, Dance, Fan, Willamowski and Bray. SLCV 2004. Visual Categorization With Bags of Keypoints.

# challenges

- thousands of local features per image
- vocabularies may need to be very large
- bag-of-words invariant but not discriminative
- spatial matching does not scale well
- quantization hurts
- burstiness of visual elements hurts
- need for efficient nearest neighbor search
- datasets are redundant

# outline – part I

2 context

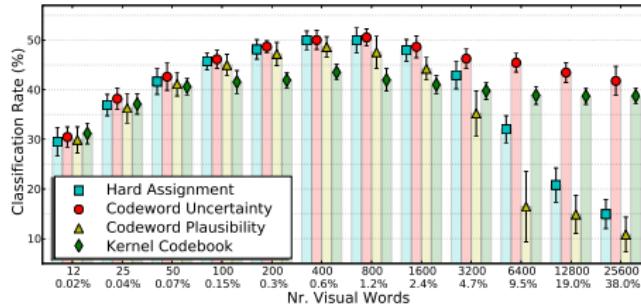
3 visual vocabularies

4 spatial matching

5 beyond vocabularies

6 exploring photo collections

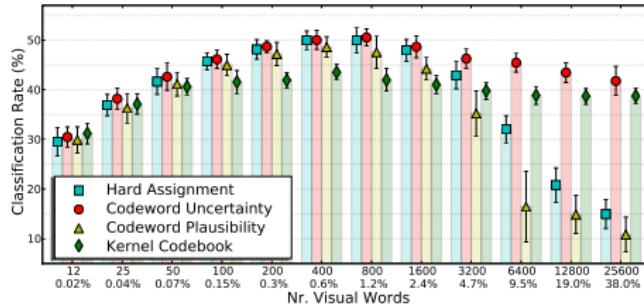
# vocabulary size



classification

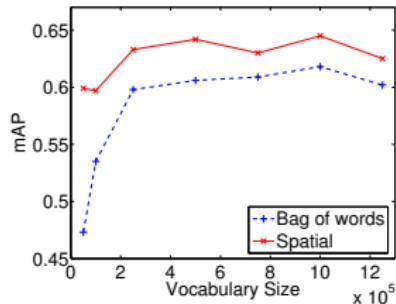
● thousands

# vocabulary size



## classification

- thousands



## instance-level retrieval

- millions

Gemert, Geusebroek, Veenman and Smeulders. ECCV 2008. Kernel Codebooks for Scene Categorization.

Philbin, Chum, Isard, Sivic and Zisserman. CVPR 2007. Object Retrieval With Large Vocabularies and Fast Spatial Matching.

# problems

- with  $k = 10^6$  visual words and  $n = 10^7$  descriptors, vocabulary learning is very **expensive**: only variants of  **$k$ -means**
- for each value of  $k$  tested, one needs to not only learn the vocabulary, but also **re-index** a very large image collection

# beyond $k$ -means

## approximate $k$ -means (AKM)

- centroids updated as in  $k$ -means
- points assigned to centroids by randomized  $k$ -d trees

## approximate Gaussian mixtures (AGM)

- keep nearest neighbors between iterations and use them to model a Gaussian mixture
- dynamically estimate  $k$  by purging overlapping components

# beyond $k$ -means

## approximate $k$ -means (AKM)

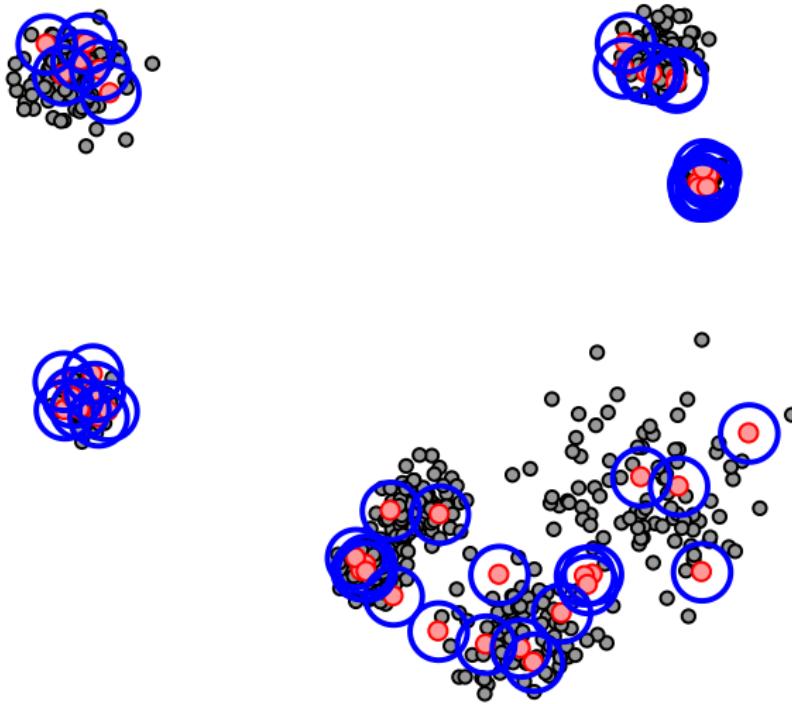
- centroids updated as in  $k$ -means
- points assigned to centroids by randomized  $k$ -d trees

## approximate Gaussian mixtures (AGM)

- keep nearest neighbors between iterations and use them to model a Gaussian mixture
- dynamically estimate  $k$  by purging overlapping components

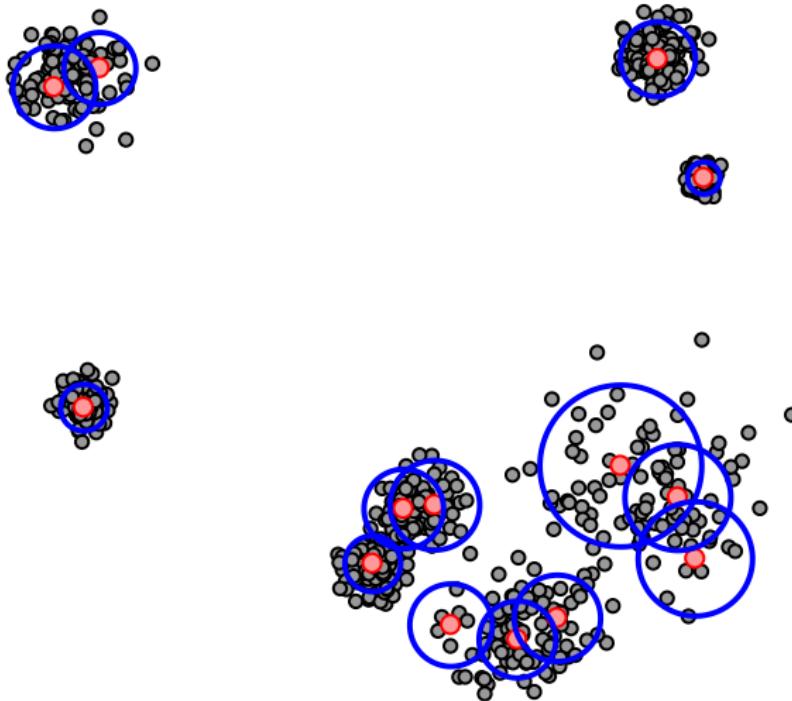
## approximate Gaussian mixtures

iteration 0: 50 clusters



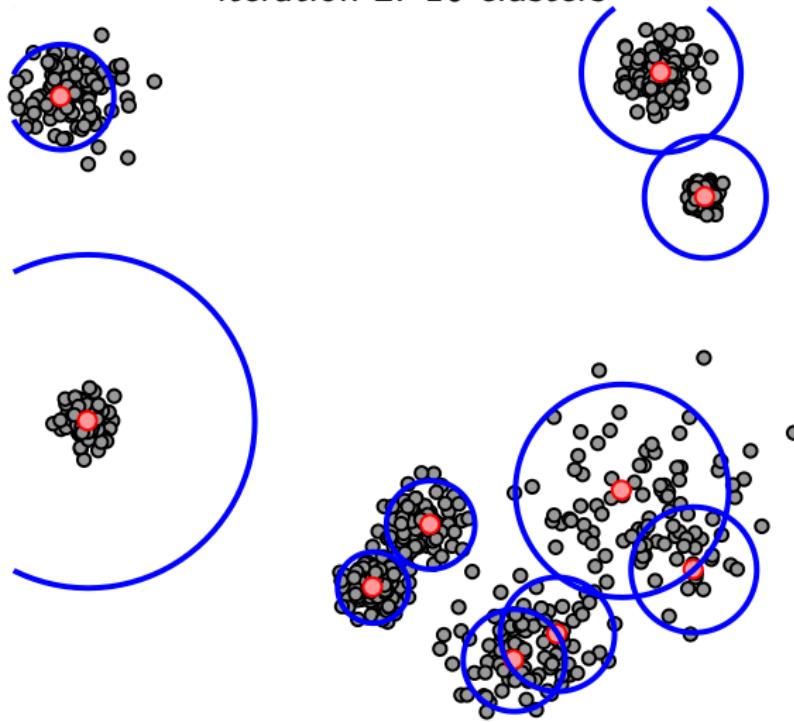
## approximate Gaussian mixtures

iteration 1: 15 clusters



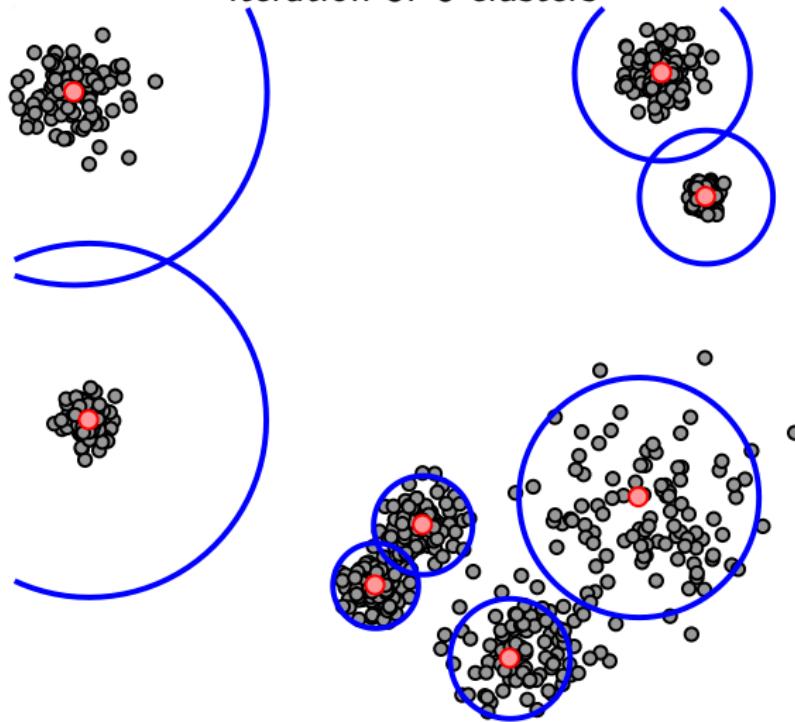
## approximate Gaussian mixtures

iteration 2: 10 clusters



## approximate Gaussian mixtures

### iteration 3: 8 clusters



# results

image search: mAP on Oxford5k

Method $k$	RAKM				AKM		AGM
	350k	500k	550k	600k	700k	550k	857k
5k	0.471	0.479	0.486	0.485	0.476	0.485	0.492
5k + 20k	0.439	0.440	0.448	0.441	0.437	0.447	0.459
5k + 1M	-	-	0.250	-	-	-	0.280

- RAKM roughly equivalent to AKM, only faster
- AGM superior, with  $k = 857k$  automatically inferred in a single run

# outline – part I

2 context

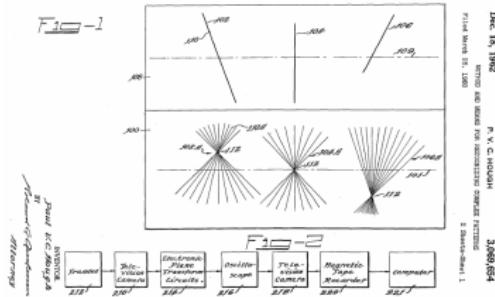
3 visual vocabularies

4 spatial matching

5 beyond vocabularies

6 exploring photo collections

# robust matching



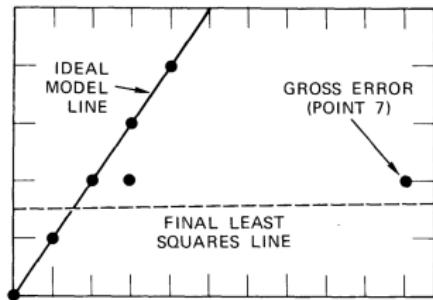
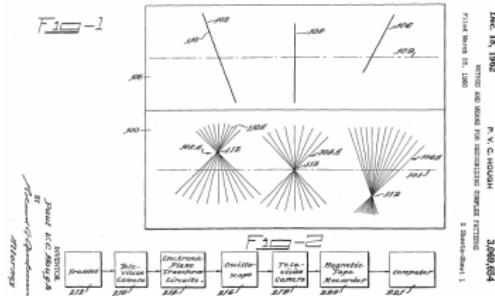
## Hough transform

- detect patterns by a voting process in parameter space

Hough. US Patent 1962. Method and Means for Recognizing Complex Patterns.

Fischler and Bolles. CACM 1981. Random Sample Consensus: A Paradigm for Model Fitting With Applications to Image Analysis and Automated Cartography.

# robust matching



## Hough transform

- detect patterns by a voting process in parameter space

## random sample consensus (RANSAC)

- iteratively generate hypotheses at random, fit model, and verify hypotheses by counting inliers

Hough. US Patent 1962. Method and Means for Recognizing Complex Patterns.

Fischler and Bolles. CACM 1981. Random Sample Consensus: A Paradigm for Model Fitting With Applications to Image Analysis and Automated Cartography.

# using local shape

a single correspondence of SIFT features yields a 4-dof transformation



## Lowe

- **hypotheses**: sparse Hough voting in 4-dimensional space
- **verification**: find inliers for bins with at least 3 votes

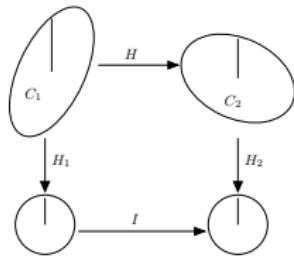
# using local shape

a single correspondence of SIFT features yields a 4-dof transformation



## Lowe

- **hypotheses**: sparse Hough voting in 4-dimensional space
- **verification**: find inliers for bins with at least 3 votes



## fast spatial matching (FSM)

- 3, 4 or 5-dof transformation
- RANSAC with one hypothesis per correspondence

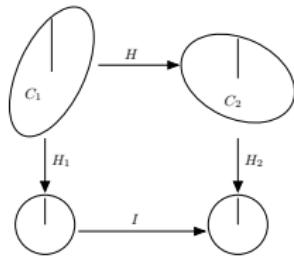
# using local shape

a single correspondence of SIFT features yields a 4-dof transformation



## Lowe

- **hypotheses**: sparse Hough voting in 4-dimensional space
- **verification**: find inliers for bins with at least 3 votes

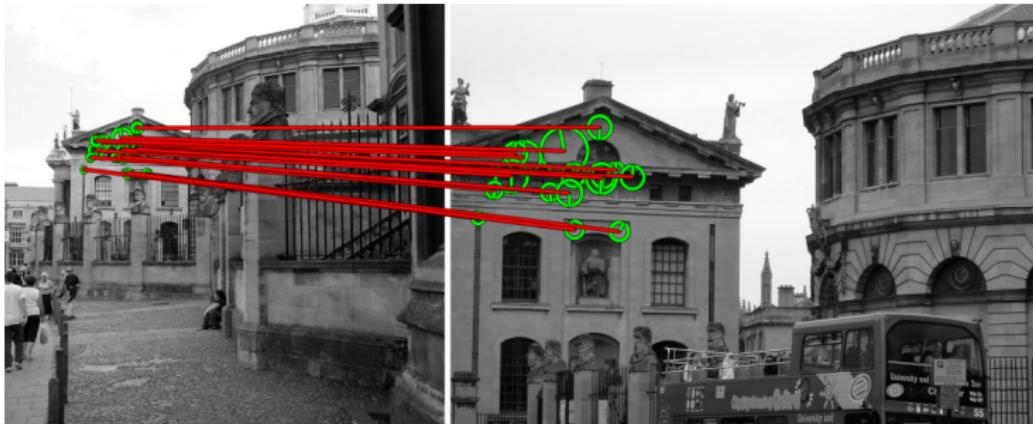


## fast spatial matching (FSM)

- 3, 4 or 5-dof transformation
- RANSAC with one hypothesis per correspondence

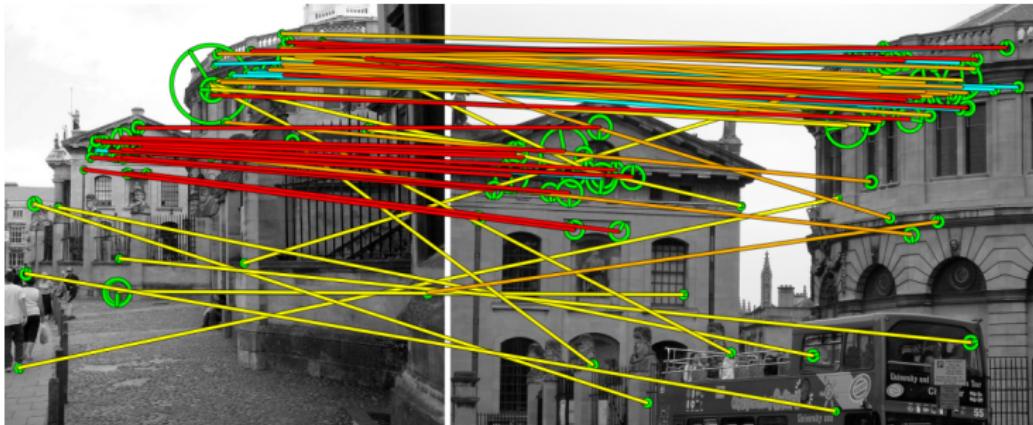
**both are quadratic in the number of correspondences**

## Hough pyramid matching (HPM)



## fast spatial matching

## Hough pyramid matching (HPM)

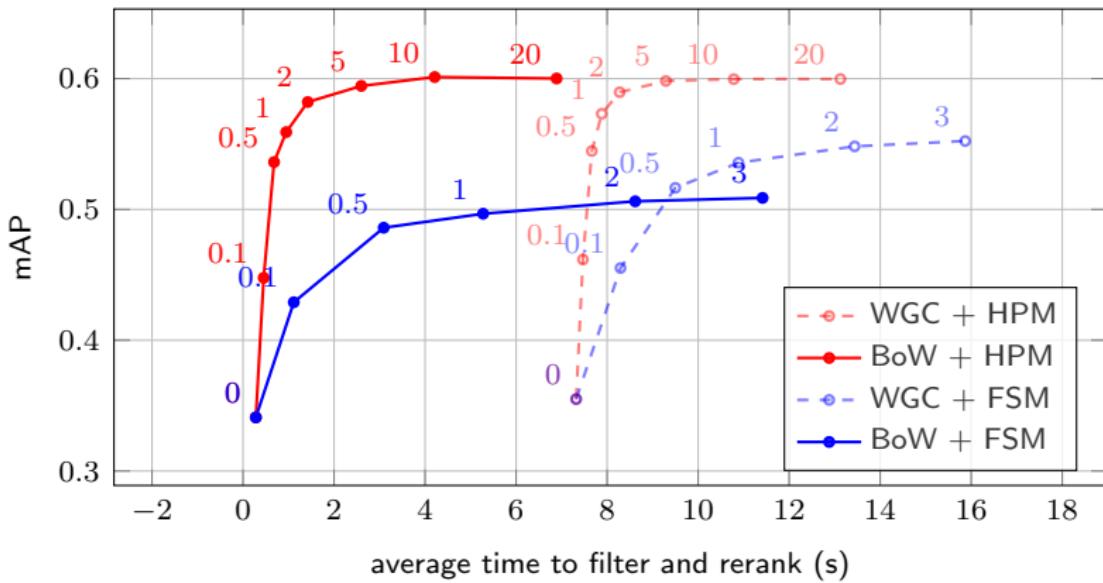


## Hough pyramid matching

- robust to deformation, **multiple surfaces**, invariant to transformations
  - **linear** in the number of correspondences; no need to count inliers

# performance vs. time

image search on World Cities 2M



- more than 10 times faster, more accurate

# outline – part I

2 context

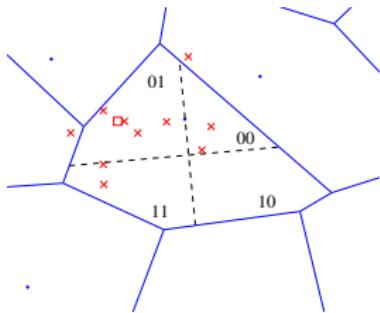
3 visual vocabularies

4 spatial matching

5 beyond vocabularies

6 exploring photo collections

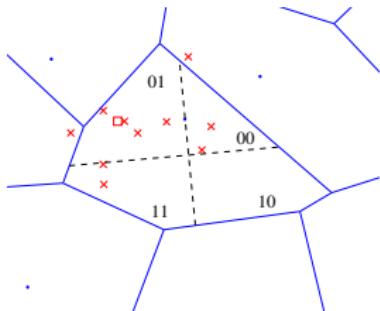
# pairwise matching vs. aggregation



## Hamming embedding (HE)

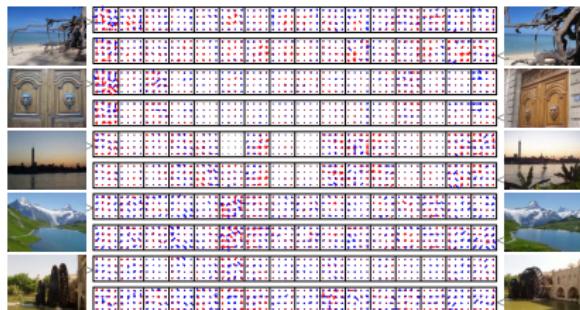
- large vocabulary
- matching of binary signatures
- **selective**: discard weak votes

# pairwise matching vs. aggregation



## Hamming embedding (HE)

- large vocabulary
- matching of binary signatures
- **selective**: discard weak votes



## vector of locally aggregated descriptors (VLAD)

- small vocabulary
- one aggregated vector per cell
- **not selective**

## aggregated selective match kernel (ASMK)

- borrow from HE the idea that descriptor pairs are **selected** by a nonlinear function

$$K_{\text{HE}}(X, Y) := \sum_{x \in X} \sum_{y \in Y} \mathbb{1}[d_{\mathbf{H}}(b(x), b(y)) \leq \tau]$$

- borrow from VLAD the idea that residuals are **aggregated** per cell

$$K_{\text{VLAD}}(X, Y) := V(X)^{\top} V(Y) = \sum_{x \in X} \sum_{y \in Y} r(x)^{\top} r(y)$$

- combine aggregation **within** cells with selectivity **between** cells

$$K_{\text{ASMK}}(X, Y) := \sigma_{\alpha}(\hat{V}(X)^{\top} \hat{V}(Y))$$

where  $\hat{x} := x / \|x\|$  and  $\sigma_{\alpha}$  a nonlinear **selectivity** function

## aggregated selective match kernel (ASMK)

- borrow from HE the idea that descriptor pairs are **selected** by a nonlinear function

$$K_{\text{HE}}(X, Y) := \sum_{x \in X} \sum_{y \in Y} \mathbb{1}[d_{\mathsf{H}}(b(x), b(y)) \leq \tau]$$

- borrow from VLAD the idea that residuals are **aggregated** per cell

$$K_{\text{VLAD}}(X, Y) := V(X)^{\top} V(Y) = \sum_{x \in X} \sum_{y \in Y} r(x)^{\top} r(y)$$

- combine aggregation **within** cells with selectivity **between** cells

$$K_{\text{ASMK}}(X, Y) := \sigma_{\alpha}(\hat{V}(X)^{\top} \hat{V}(Y))$$

where  $\hat{x} := x / \|x\|$  and  $\sigma_{\alpha}$  a nonlinear **selectivity** function

## aggregated selective match kernel (ASMK)

- borrow from HE the idea that descriptor pairs are **selected** by a nonlinear function

$$K_{\text{HE}}(X, Y) := \sum_{x \in X} \sum_{y \in Y} \mathbb{1}[d_{\mathcal{H}}(b(x), b(y)) \leq \tau]$$

- borrow from VLAD the idea that residuals are **aggregated** per cell

$$K_{\text{VLAD}}(X, Y) := V(X)^{\top} V(Y) = \sum_{x \in X} \sum_{y \in Y} r(x)^{\top} r(y)$$

- combine aggregation **within** cells with selectivity **between** cells

$$K_{\text{ASMK}}(X, Y) := \sigma_{\alpha}(\hat{V}(X)^{\top} \hat{V}(Y))$$

where  $\hat{x} := x / \|x\|$  and  $\sigma_{\alpha}$  a nonlinear **selectivity** function

# impact of selectivity

$$\alpha = 3, \tau = 0.0$$



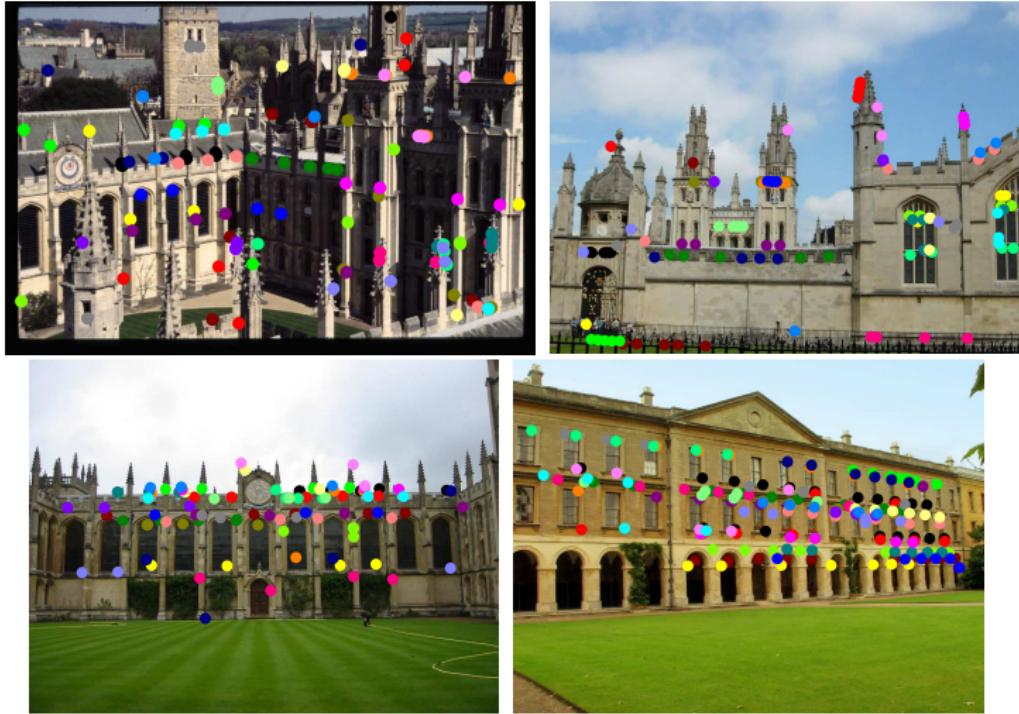
$$\alpha = 3, \tau = 0.25$$



correspondences weighed based on confidence

# impact of aggregation and burstiness

$k = 65k$  as in HE



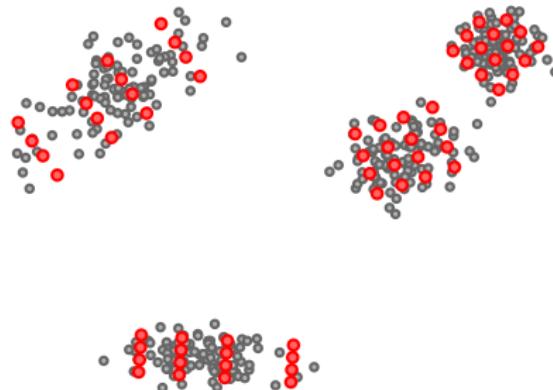
# results

## image search: mAP

Dataset	MA	Oxf5k	Oxf105k	Par6k	Holiday
ASMK*		76.4	69.2	74.4	80.0
ASMK*	✓	80.4	75.0	77.0	81.0
ASMK		78.1	-	76.0	81.2
ASMK	✓	81.7	-	78.2	82.2
HE [Jégou <i>et al.</i> '10]		51.7	-	-	74.5
HE [Jégou <i>et al.</i> '10]	✓	56.1	-	-	77.5
HE-BURST [Jain <i>et al.</i> '10]		64.5	-	-	78.0
HE-BURST [Jain <i>et al.</i> '10]	✓	67.4	-	-	79.6
Fine vocab. [Mikulík <i>et al.</i> '10]	✓	74.2	67.4	74.9	74.9

- last state of the art before deep learning
- still state of the art on CNN features

# locally optimized product quantization



- builds on PQ, searching fast in the compressed domain
- better captures the support of data distribution
- state of the art at billion scale for years
- deployed on entire Flickr collection

Jégou, Douze and Schmid. PAMI 2011. Product Quantization for Nearest Neighbor Search.

Kalantidis and Avrithis. CVPR 2014. Locally Optimized Product Quantization for Approximate Nearest Neighbor Search.

# outline – part I

2 context

3 visual vocabularies

4 spatial matching

5 beyond vocabularies

6 exploring photo collections

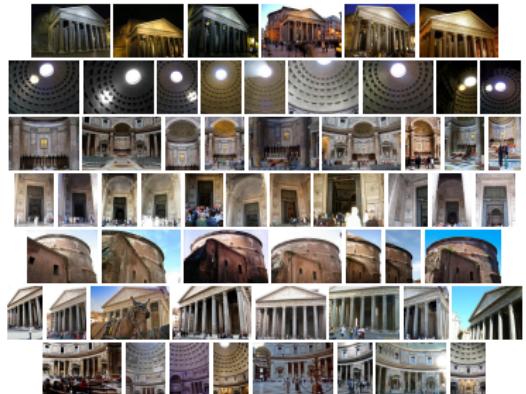
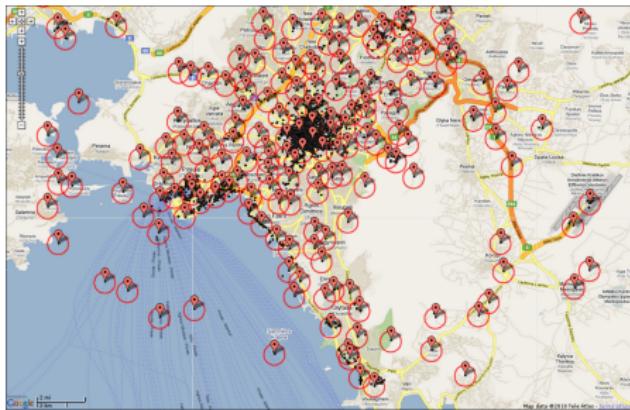
## community photo collections

- applications: browsing, 3D reconstruction, location/landmark recognition
  - focus on **popular** subsets like landmarks and points of interest



# view clustering

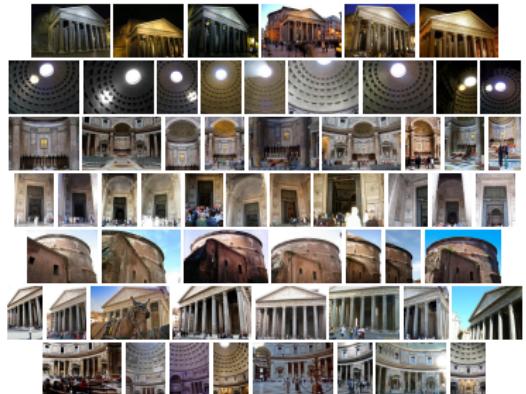
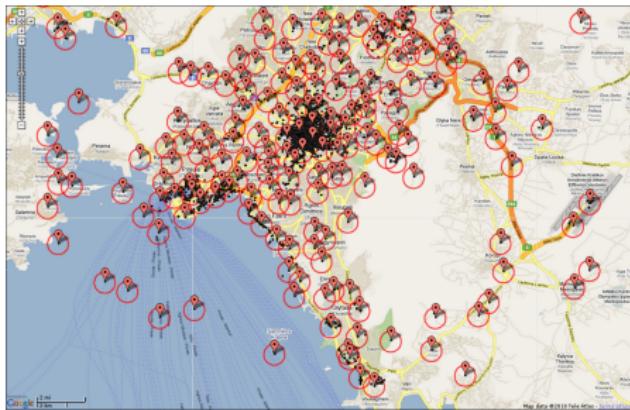
- geo clustering: according to geographic location
- visual clustering: according to visual similarity (inliers)



- both landmark and non-landmark images

# view clustering

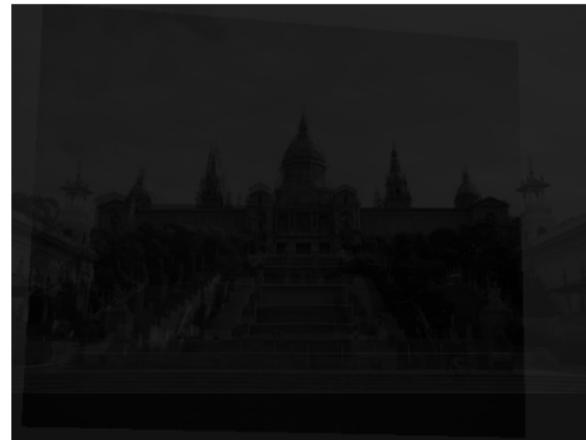
- geo clustering: according to geographic location
- visual clustering: according to visual similarity (inliers)



- both **landmark** and **non-landmark** images

# view alignment

aligned images



Avrithis, Kalantidis, Tolias and Spyrou. ACM-MM 2010. Retrieving Landmark and Non-Landmark Images From Community Photo Collections.

# view alignment

## aligned images



# view alignment

## aligned images



# view alignment

## aligned images



# view alignment

## aligned images



# view alignment

## aligned images



# view alignment

## aligned images



Avrithis, Kalantidis, Tolias and Spyrou. ACM-MM 2010. Retrieving Landmark and Non-Landmark Images From Community Photo Collections.

# view alignment

## aligned images



Avrithis, Kalantidis, Tolias and Spyrou. ACM-MM 2010. Retrieving Landmark and Non-Landmark Images From Community Photo Collections.

# view alignment

## aligned images



Avrithis, Kalantidis, Tolias and Spyrou. ACM-MM 2010. Retrieving Landmark and Non-Landmark Images From Community Photo Collections.

# view alignment

## aligned images



Avrithis, Kalantidis, Tolias and Spyrou. ACM-MM 2010. Retrieving Landmark and Non-Landmark Images From Community Photo Collections.

# view alignment

## aligned images



Avrithis, Kalantidis, Tolias and Spyrou. ACM-MM 2010. Retrieving Landmark and Non-Landmark Images From Community Photo Collections.

# view alignment

## aligned images



Avrithis, Kalantidis, Tolias and Spyrou. ACM-MM 2010. Retrieving Landmark and Non-Landmark Images From Community Photo Collections.

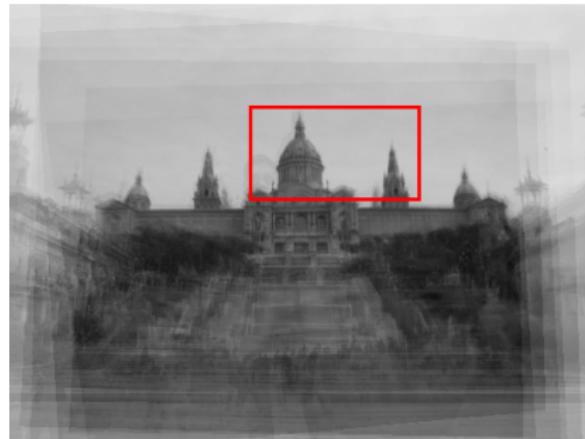
# scene map construction

## before feature clustering



# scene map construction

after feature clustering



Avrithis, Kalantidis, Tolias and Spyrou. ACM-MM 2010. Retrieving Landmark and Non-Landmark Images From Community Photo Collections.

# results

## image search on European Cities 1M

Method	Time	mAP
Baseline BoW	1.03s	0.642
QE <sub>1</sub>	20.30s	0.813
QE <sub>2</sub>	2.51s	0.686
Scene maps	1.29s	0.824

- QE<sub>1</sub>: iterative query expansion, re-query using the retrieved images and merge, 3 times iteratively
- QE<sub>2</sub>: create scene map using the initial results and re-query once
- scene maps: similar to QE<sub>1</sub> but as fast as baseline

Chum, Philbin, Sivic, Isard and Zisserman. ICCV 2007. Total Recall: Automatic Query Expansion With a Generative Feature Model for Object Retrieval.

Avrithis, Kalantidis, Tolias and Spyrou. ACM-MM 2010. Retrieving Landmark and Non-Landmark Images From Community Photo Collections.

<http://viral.image.ntua.gr>

online since 2008

# query



## results



**Suggested tags:** Buxton Memorial Fountain, Victoria Tower Gardens, London  
**Frequent user tags:** Victoria Tower Gardens, Buxton Memorial Fountain, Winchester Palace, Architecture, Victorian gothic

## Similar Images



Similarity: 0.619



Similarity: 0.491



Similarity: 0.397



Similarity: 0.385  
Details Original

# suggested tags



**Suggested tags:** Buxton Memorial Fountain, Victoria Tower Gardens, London

**Frequent user tags:** Victoria Tower Gardens, Buxton Memorial Fountain, Winchester Palace, Architecture, Victorian gothic

## related wikipedia articles



WIKIPEDIA  
The Free Encyclopedia

- [Main page](#)
- [Contents](#)
- [Featured content](#)
- [Current events](#)
- [Random article](#)
- [Donate](#)
- 
- [Interaction](#)
- [About Wikipedia](#)
- [Community portal](#)
- [Recent changes](#)
- [Contact Wikipedia](#)
- [Help](#)

- Toolbox
  - What links here
  - Related changes
  - Upload file
  - Special pages
  - Permanent link
  - Cite this page

▼ Print\export

## Article Discussion

Read Edit View history

New features  Log in / create account

Victoria Tower Gardens

From Wikipedia, the free encyclopedia

**Victoria Tower Gardens** is a public park along the north bank of the River Thames in London. As its name suggests, it is adjacent to the Victoria Tower, the south-western corner of the Palace of Westminster. The park, which extends southwards from the Palace to Lambeth Bridge, sandwiched between Millbank and the river, also forms part of the **Thames Embankment**.

**Contents** [hide]

- [1 Features](#)
- [2 Transport](#)
- [3 History](#)
- [4 External links](#)
- [5 References](#)

## Features

100

### The park features

- A reproduction of the sculpture *The Burghers of Calais* by Auguste Rodin, purchased by the British Government in 1911 and positioned in the Gardens in 1915.
  - A 1930 statue of the suffragette Emmeline Pankhurst, by A.G. Walker.
  - The Buxton Memorial Fountain – originally constructed in Parliament Square, this was removed in 1940 and placed in its present position in 1957. It was commissioned by Charles Buxton MP to commemorate the emancipation of slaves in 1834, dedicated to his father Thomas Fowell Buxton, and designed by Gothic architect Samuel Sanders Teulon (1812–1873) in 1865.
  - A stone wall with two modern-style goats with kids – situated at the southern end of the Gardens.

## Transport

[edit]



Victoria Tower Gardens, 2005, with the Buxton Memorial Fountain at the front and the Palace of Westminster in the background.

# VIRaL Explore



Kalantidis, Tolias, Avrithis, Phinikettos, Spyrou, Mylonas and Kollias. MTAP 2011. VIRaL: Visual Image Retrieval and Localization.

# VIRaL Explore



# VIRaL Routes

**Identified landmarks**

**Frequent user tags**

**User images**

**Similar images**

**Viewing Venice by ykalant.**

**Change photo set**

## achievements and more challenges

- one-off construction of vocabularies
- fast and more accurate spatial matching
- beyond BoW: approximate descriptors, fighting burstiness
- nearest neighbor search in compressed domain
- dataset-wide analysis improves image representation
- widespread dissemination of novel applications
- either high quality or compact representation

## achievements and more challenges

- one-off construction of vocabularies
- fast and more accurate spatial matching
- beyond BoW: approximate descriptors, fighting burstiness
- nearest neighbor search in compressed domain
- dataset-wide analysis improves image representation
- widespread dissemination of novel applications
- either high quality or compact representation

part II

exploring deeper

# outline – part II

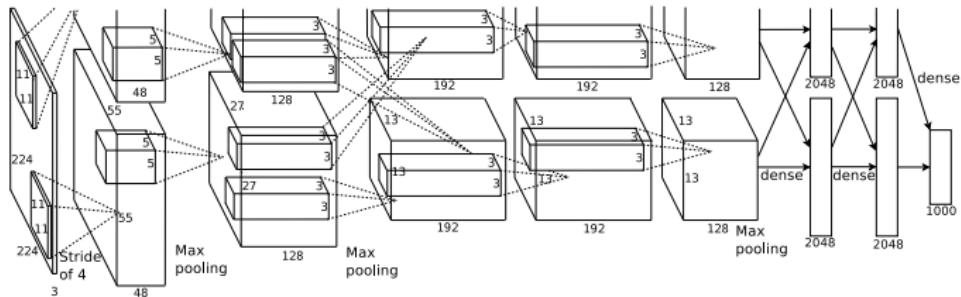
7 context

8 searching on manifolds

9 spatial matching

10 discovering objects

# AlexNet



learning visual representations from raw data works at scale

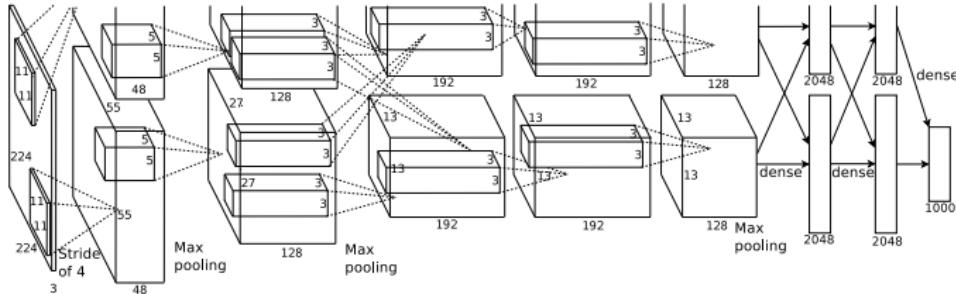
## CNN, SGD backprop

ImageNet  
(1.2M images)

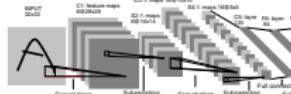
## graphics processing units (GPU)

rectified linear  
unit (ReLU)

# AlexNet



learning visual representations from raw data works at scale



CNN, SGD  
backprop

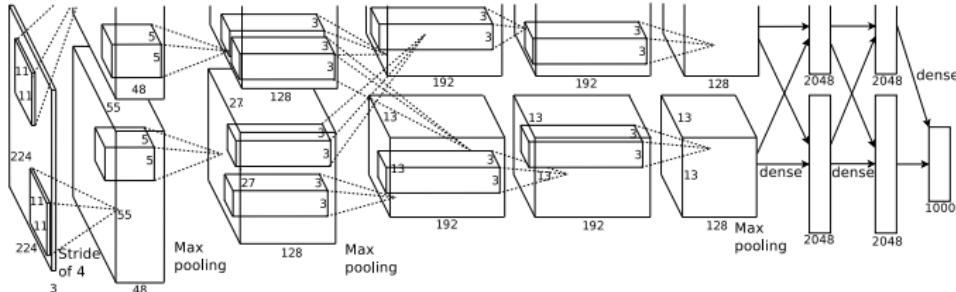
ImageNet  
(1.2M images)

graphics processing  
units (GPU)

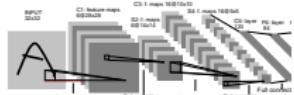
rectified linear  
unit (ReLU)

LeCun, Boser, Denker et al . NIPS 1990. Handwritten Digit Recognition with a Back-Propagation Network.  
Krizhevsky, Sutskever and Hinton. NIPS 2012. ImageNet Classification with Deep Convolutional Neural Networks.

# AlexNet



learning visual representations from raw data works at scale



CNN, SGD  
backprop



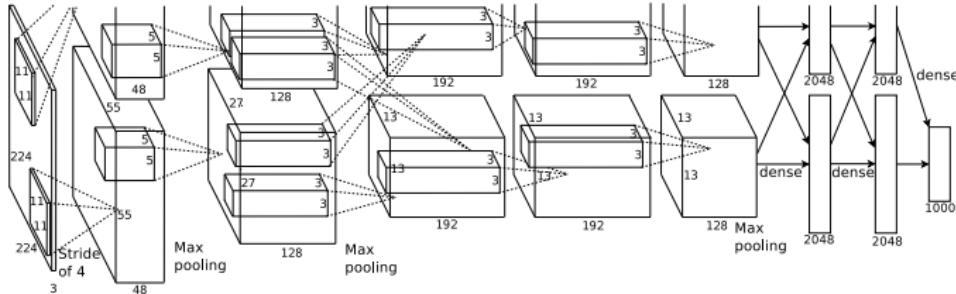
ImageNet  
(1.2M images)

graphics processing  
units (GPU)

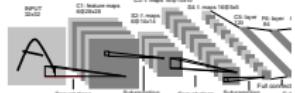
rectified linear  
unit (ReLU)

Russakovsky, Deng, Su, Krause et al. 2014. Imagenet Large Scale Visual Recognition Challenge.  
Krizhevsky, Sutskever and Hinton. NIPS 2012. ImageNet Classification with Deep Convolutional Neural Networks.

# AlexNet



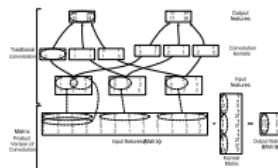
learning visual representations from raw data works at scale



CNN, SGD  
backprop



ImageNet  
(1.2M images)

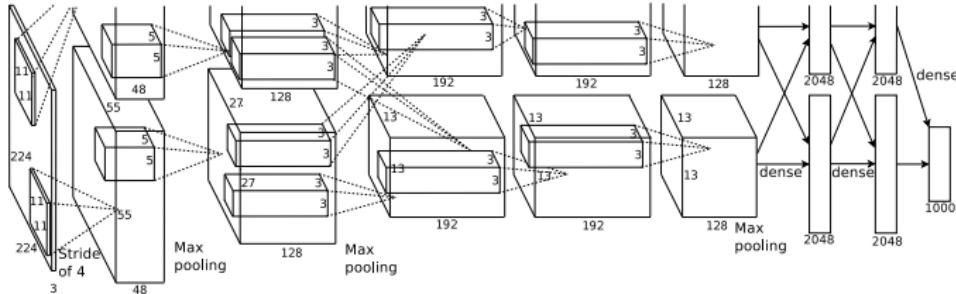


graphics processing  
units (GPU)

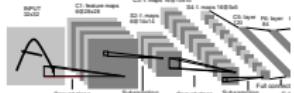
rectified linear  
unit (ReLU)

Chellapilla, Puri and Simard. FHR 2006. High Performance Convolutional Neural Networks for Document Processing.  
Krizhevsky, Sutskever and Hinton. NIPS 2012. ImageNet Classification with Deep Convolutional Neural Networks.

# AlexNet



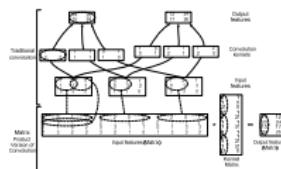
learning visual representations from raw data works at scale



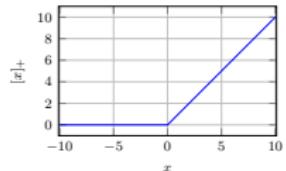
CNN, SGD  
backprop



ImageNet  
(1.2M images)



graphics processing  
units (GPU)

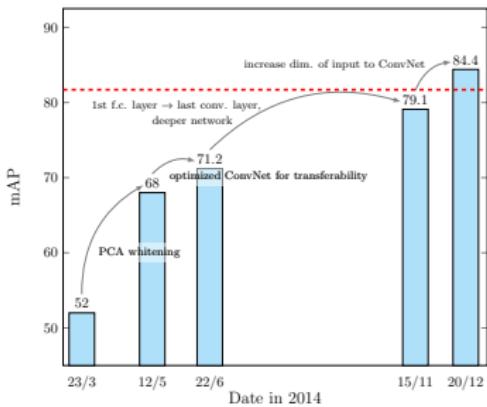


rectified linear  
unit (ReLU)

Nair and Hinton. ICML 2010. Rectified Linear Units Improve Restricted Boltzmann Machines.

Krizhevsky, Sutskever and Hinton. NIPS 2012. ImageNet Classification with Deep Convolutional Neural Networks.

# instance-level tasks



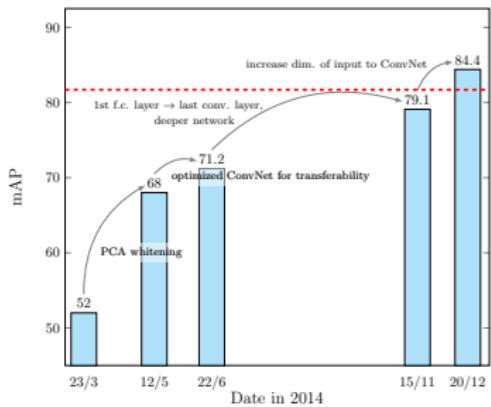
## regional CNN features

- jump more than 30% mAP in few months
- outperform SIFT pipeline

Razavian, Sullivan, Maki and Carlsson. arXiv 2015. Visual Instance Retrieval with Deep Convolutional Networks.

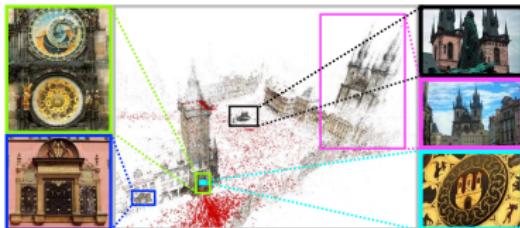
Radenovic, Tolias, Chum. ECCV 2016. CNN Image Retrieval Learns From BoW: Unsupervised Fine-Tuning with Hard Examples.

# instance-level tasks



## regional CNN features

- jump more than 30% mAP in few months
- outperform SIFT pipeline



## self-supervision

- max-pooling (MAC/R-MAC), generalized mean (GeM)
- SfM pipeline based on SIFT, BoW and RANSAC

## opportunities and challenges

- powerful global representation
- feature space still exhibits manifold structure
- graph-based methods now feasible but still do not scale well
- regional or local information often overlooked
- richness of convolutional activations not well understood
- dataset-wide analysis often missing in favor of stochastic updates

# outline – part II

7 context

8 searching on manifolds

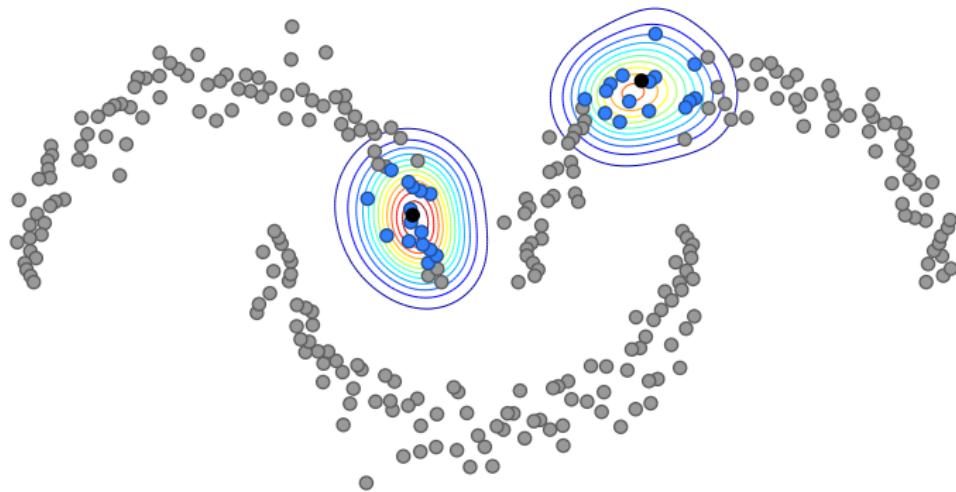
9 spatial matching

10 discovering objects

## graph-based methods

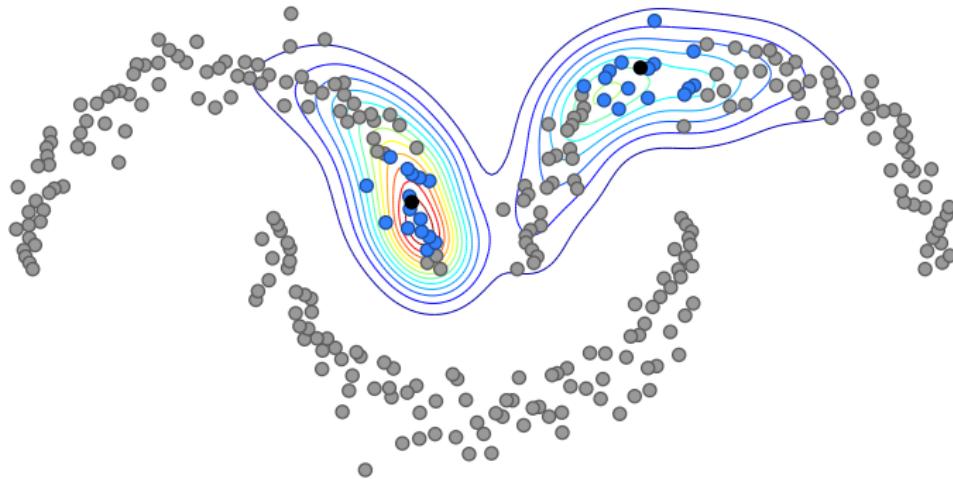
now that a high-quality representation is possible with just one or few vectors per image, graph-based methods are more relevant than ever

## ranking on manifolds (diffusion)



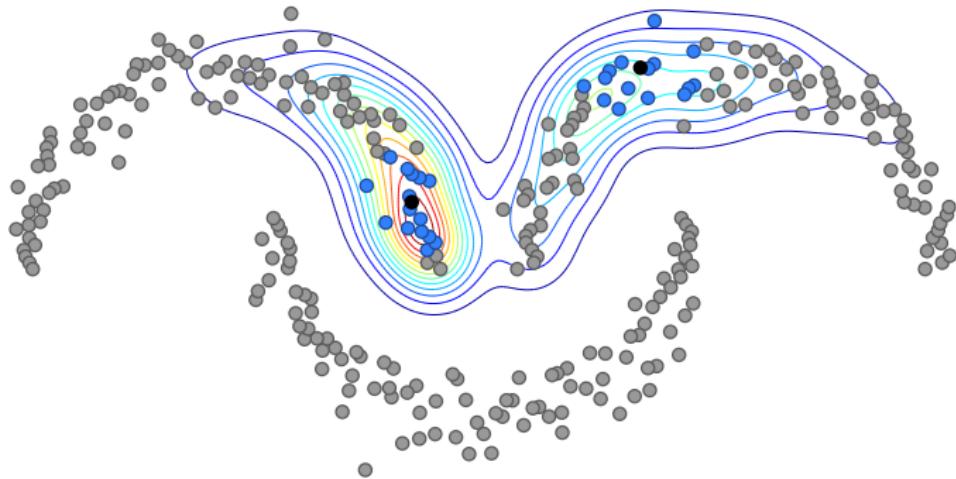
- data points ( $\circ$ ), query points ( $\bullet$ ), nearest neighbors ( $\oplus$ )
  - iteration 0  $\times$  30

## ranking on manifolds (diffusion)



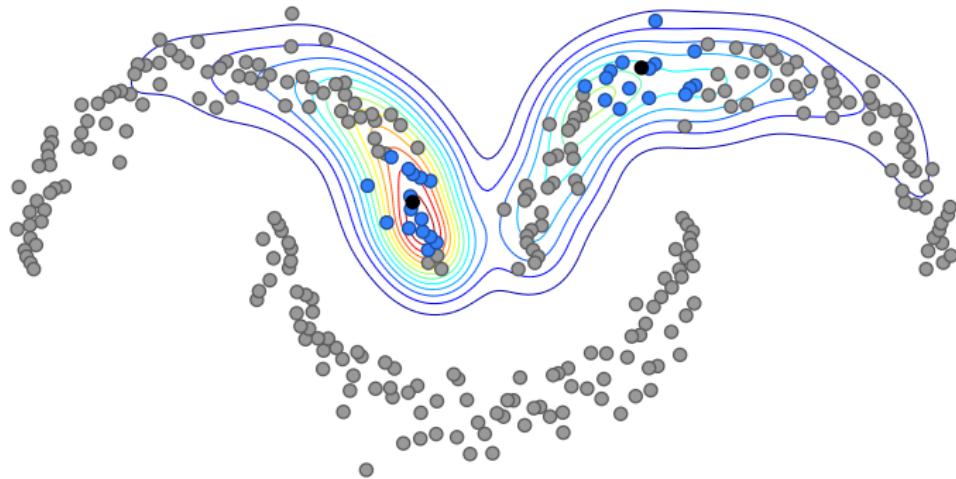
- data points ( $\circ$ ), query points ( $\bullet$ ), nearest neighbors ( $\circledast$ )
- iteration  $1 \times 30$

## ranking on manifolds (diffusion)



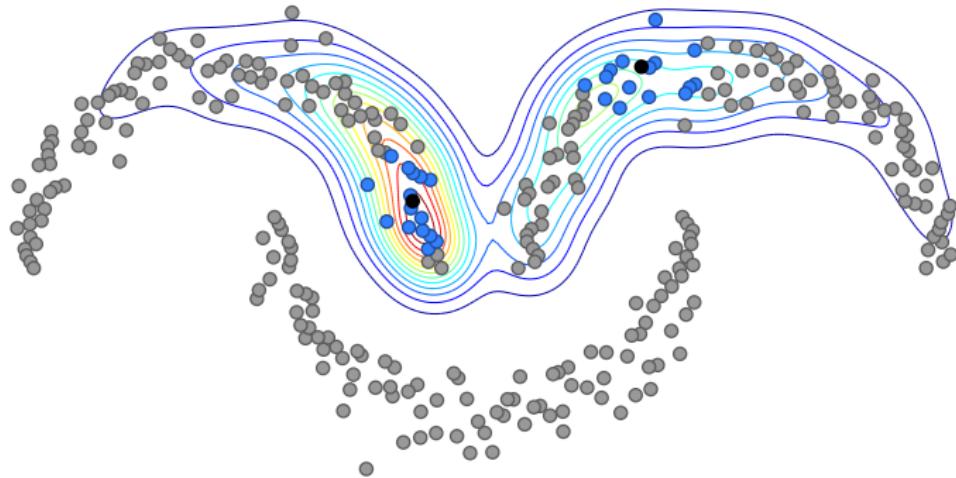
- data points (•), query points (•), nearest neighbors (•)
- iteration  $2 \times 30$

## ranking on manifolds (diffusion)



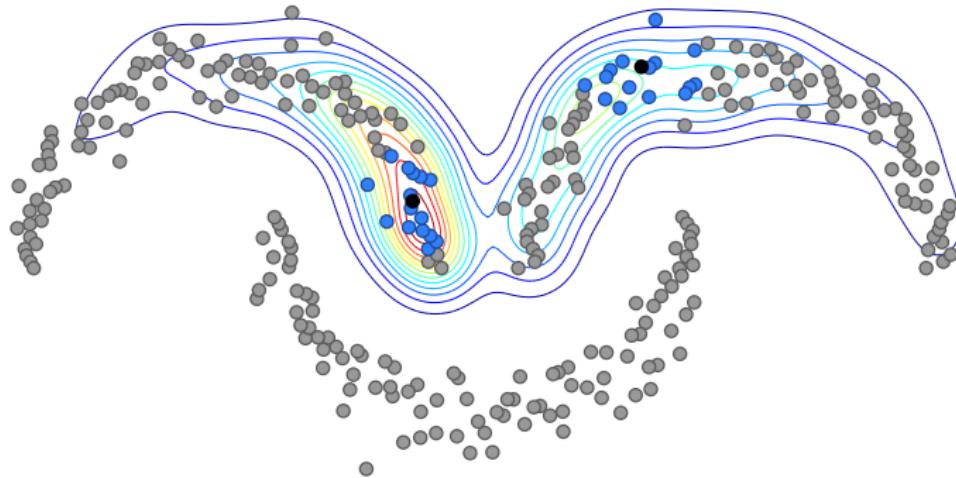
- data points ( $\circ$ ), query points ( $\bullet$ ), nearest neighbors ( $\oplus$ )
  - iteration 3  $\times$  30

## ranking on manifolds (diffusion)



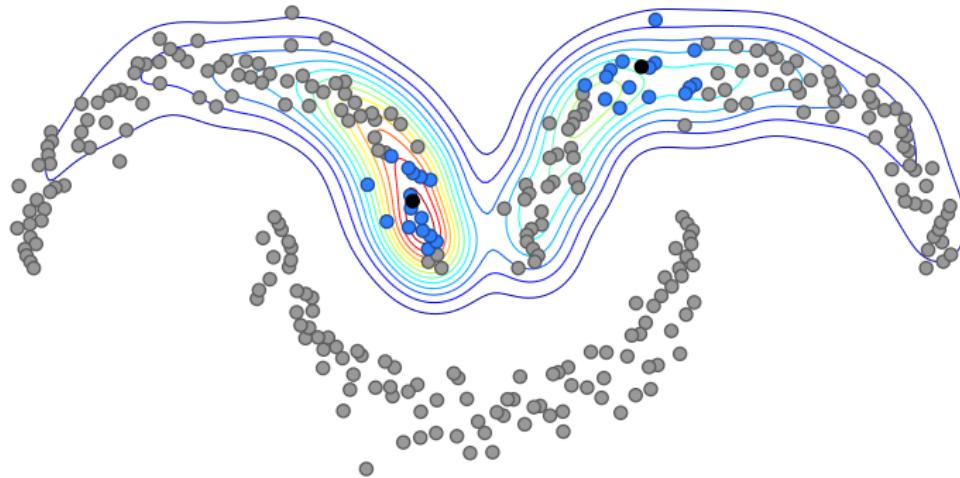
- data points ( $\circ$ ), query points ( $\bullet$ ), nearest neighbors ( $\oplus$ )
  - iteration 4  $\times$  30

## ranking on manifolds (diffusion)



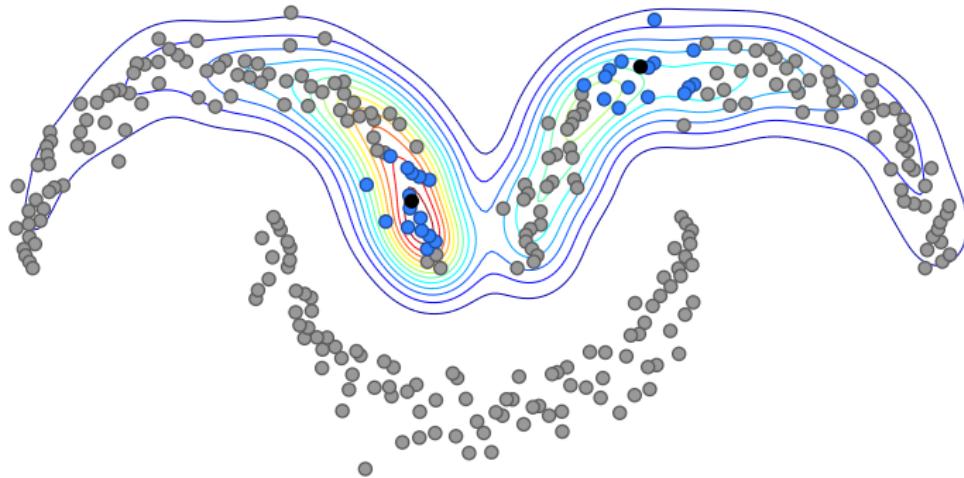
- data points ( $\circ$ ), query points ( $\bullet$ ), nearest neighbors ( $\oplus$ )
  - iteration 5  $\times$  30

## ranking on manifolds (diffusion)



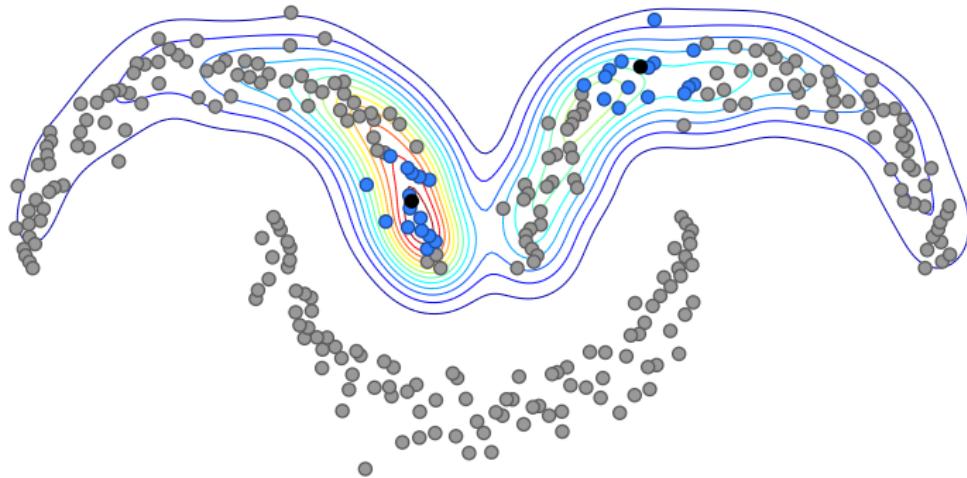
- data points ( $\circ$ ), query points ( $\bullet$ ), nearest neighbors ( $\oplus$ )
  - iteration 6  $\times$  30

## ranking on manifolds (diffusion)



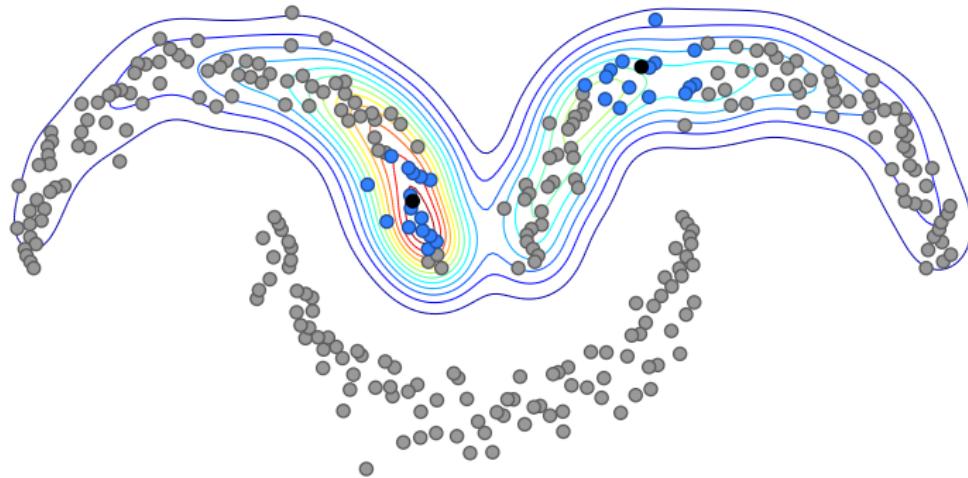
- data points (•), query points (•), nearest neighbors (•)
- iteration  $7 \times 30$

## ranking on manifolds (diffusion)



- data points ( $\circ$ ), query points ( $\bullet$ ), nearest neighbors ( $\oplus$ )
- iteration  $8 \times 30$

## ranking on manifolds (diffusion)



- data points ( $\circ$ ), query points ( $\bullet$ ), nearest neighbors ( $\oplus$ )
  - iteration  $9 \times 30$

# ranking on manifolds (diffusion)

- random walk with restart (RWR)

$$\mathbf{f}^{(\tau)} := \alpha \mathcal{W} \mathbf{f}^{(\tau-1)} + (1 - \alpha) \mathbf{y}$$

where  $\mathbf{y}$ : query vector,  $\mathcal{W}$ : adjacency matrix,  $\mathbf{f}$ : ranking vector

- apply to regional CNN features
- solve linear system

$$\mathcal{L}_\alpha \mathbf{f} = \mathbf{y}$$

by conjugate gradient (CG) method, where regularized Laplacian

$$\mathcal{L}_\alpha := \frac{I - \alpha \mathcal{W}}{1 - \alpha}$$

Zhou, Weston, Gretton, Bousquet and Schölkopf. NIPS 2003. Ranking on Data Manifolds.

Iscen, Tolias, Avrithis, Furun and Chum. CVPR 2017. Efficient Diffusion on Region Manifolds: Recovering Small Objects With Compact CNN Representations.

# ranking on manifolds (diffusion)

- random walk with restart (RWR)

$$\mathbf{f}^{(\tau)} := \alpha \mathcal{W} \mathbf{f}^{(\tau-1)} + (1 - \alpha) \mathbf{y}$$

where  $\mathbf{y}$ : query vector,  $\mathcal{W}$ : adjacency matrix,  $\mathbf{f}$ : ranking vector

- apply to regional CNN features
- solve linear system

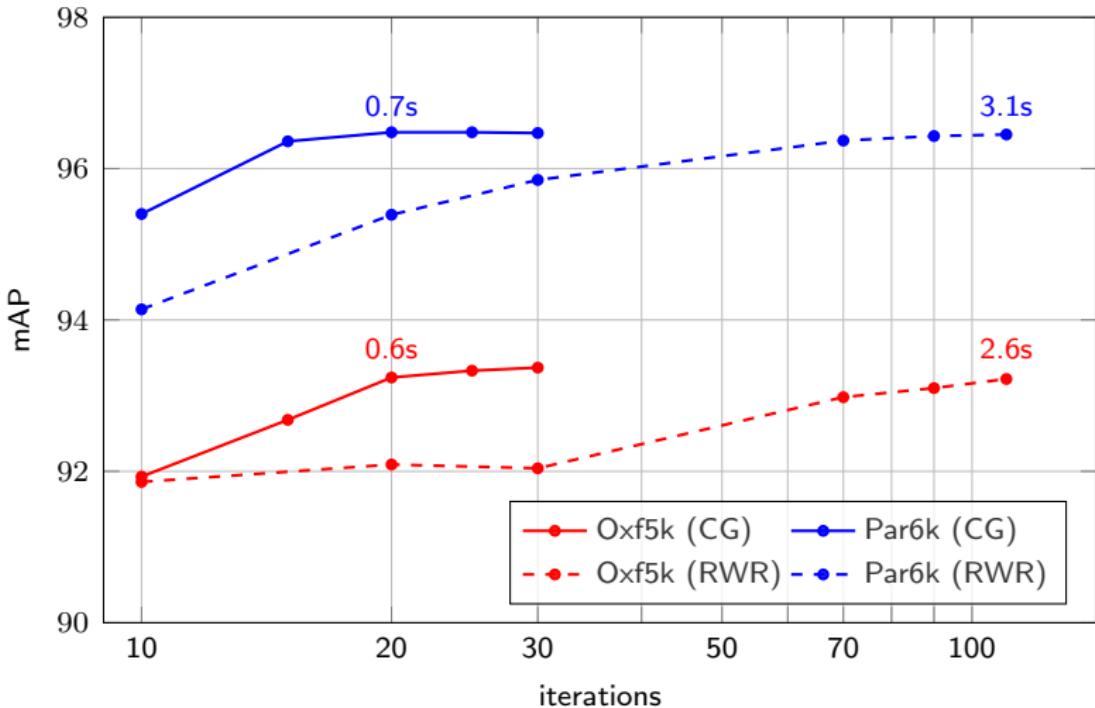
$$\mathcal{L}_\alpha \mathbf{f} = \mathbf{y}$$

by conjugate gradient (CG) method, where regularized Laplacian

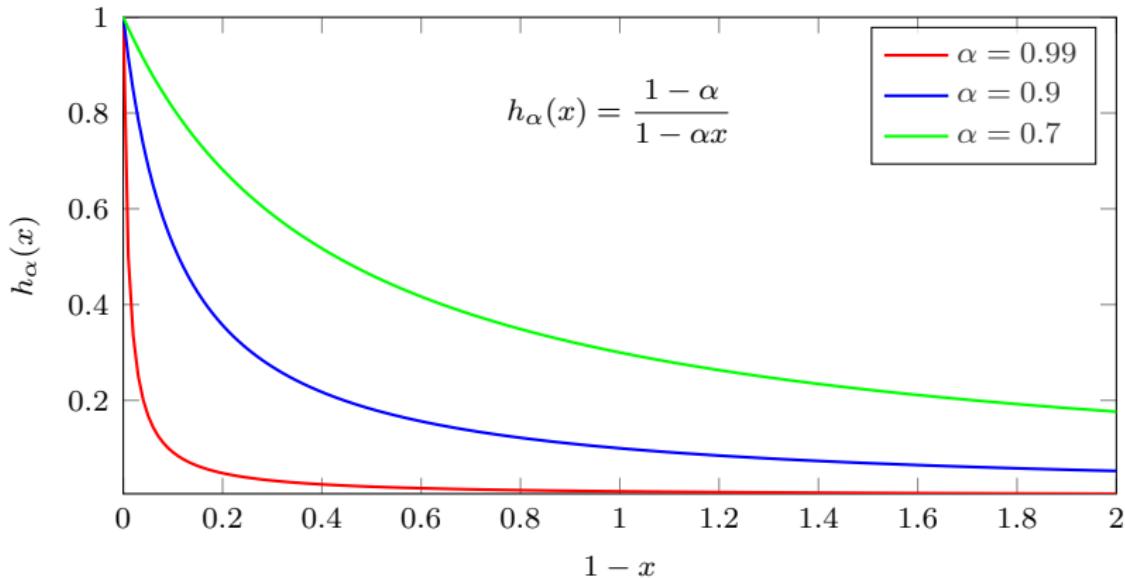
$$\mathcal{L}_\alpha := \frac{I - \alpha \mathcal{W}}{1 - \alpha}$$

## CG vs. RWR

## image search with regional VGG features ( $d = 512$ )



# fast spectral ranking (FSR)



- low-pass filtering in the frequency domain
- or, “soft” dimensionality reduction

# results

mAP using ResNet-101 features ( $d = 2,048$ )

Method	$m$	Instre	Oxf5k	Oxf105k	Par6k	Par106k
Regional Features: R-Match						
Euclidean	21	71.0	88.1	85.7	94.9	91.3
AQE	21	77.1	91.0	89.6	95.5	92.5
CG	5	88.4	95.0	90.0	96.4	95.8
FSR	5	88.5	95.1	93.0	96.5	95.2

- helps particularly on Instre, which contains small objects on background clutter
- FSR (rank  $r = 5k$ ) has same performance as CG, is **two orders of magnitude faster**, needs  $3\times$  space

# hard examples?



(AP: 92.1)



#5



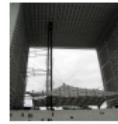
#32



#51



#70



#71



#76



#79



#126



(AP: 92.7)



#2



#4



#8



#61



#68



#72



#75



#108

- red: drift
- blue: incorrect annotations

# Oxford and Paris revisited (RevOP)



fixed annotation errors



## 1 million hard distractors



## new queries

# outline – part II

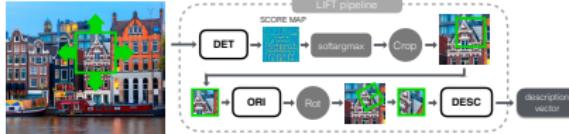
7 context

8 searching on manifolds

9 spatial matching

10 discovering objects

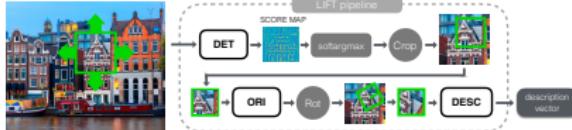
# revival of local features



## learned invariant feature transform (LIFT)

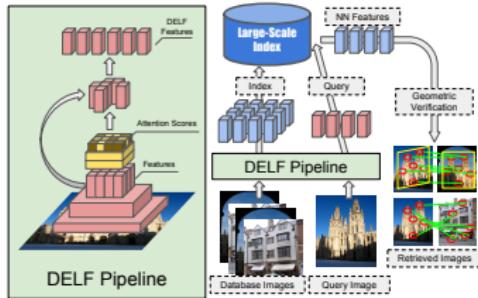
- learned SIFT: detection, orientation estimation, descriptor extraction
- trained on patch-level labels

# revival of local features



## learned invariant feature transform (LIFT)

- learned SIFT: detection, orientation estimation, descriptor extraction
- trained on patch-level labels



## deep local features (DELF)

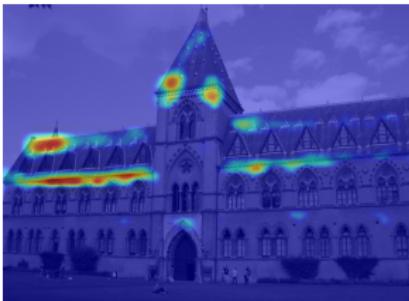
- self-attention to detect keypoints
- trained on image-level labels

Yi, Trulls, Lepetit and Fua. ECCV 2016. LIFT. Learned Invariant Feature Transform.

Noh, Araujo, Sim, Weyand and Han. ICCV 2017. Large-Scale Image Retrieval With Attentive Deep Local Features.

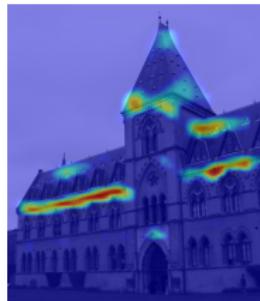
## motivation

view 1

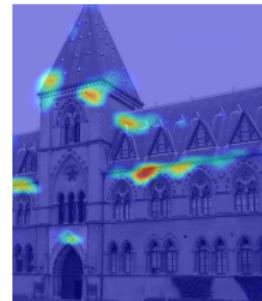


map 1

view 2



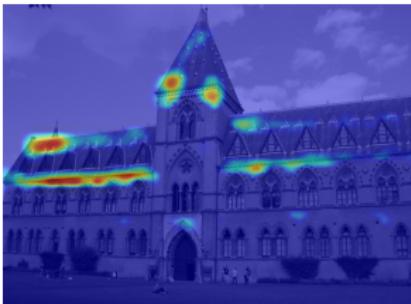
view 3



map 2

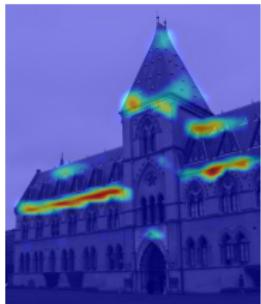
# motivation

view 1

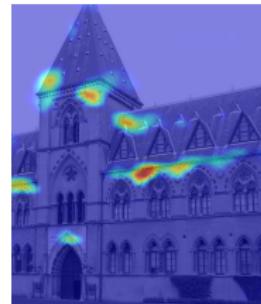


map 1

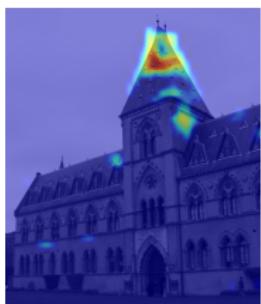
view 2



view 3



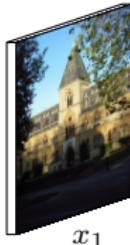
map 2



- different local features present in each feature map (channel)

# deep spatial matching (DSM)

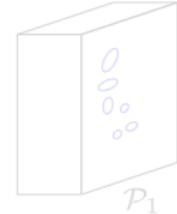
input image



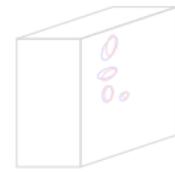
feature map



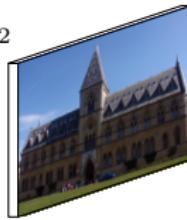
local features



inliers



$x_2$

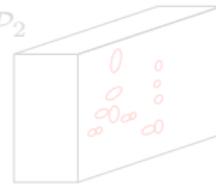


$A_2$



match  $s$

$P_2$

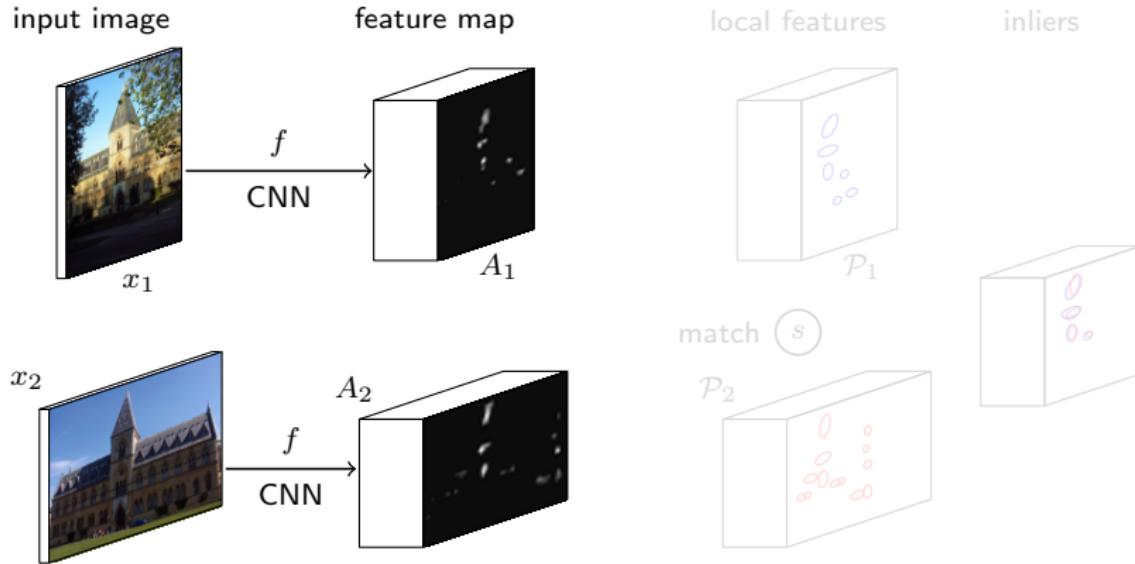


- local features detected by MSER independently per channel
- inliers found by fast spatial matching

Philbin, Chum, Isard, Sivic and Zisserman. CVPR 2007. Object Retrieval With Large Vocabularies and Fast Spatial Matching.  
Matas, Chum, Urban and Pajdla. BMVC 2002. Robust Wide Baseline Stereo From Maximally Stable Extremal Regions.

Siméoni, Avrithis and Chum. CVPR 2019. Local Features and Visual Words Emerge in Activations.

# deep spatial matching (DSM)

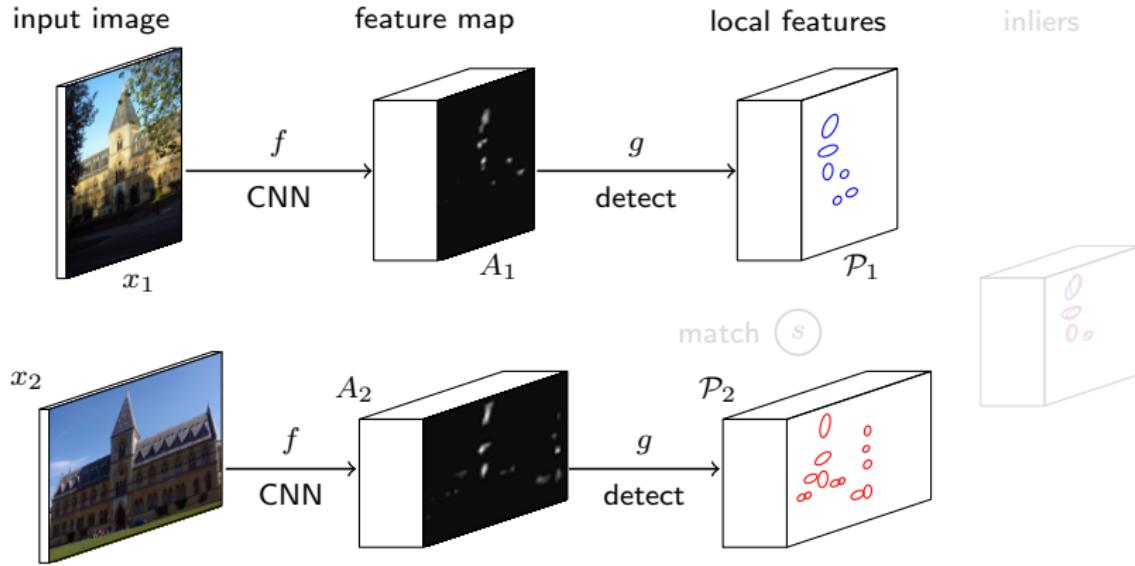


- local features detected by MSER independently per channel
- inliers found by fast spatial matching

Philbin, Chum, Isard, Sivic and Zisserman. CVPR 2007. Object Retrieval With Large Vocabularies and Fast Spatial Matching.  
Matas, Chum, Urban and Pajdla. BMVC 2002. Robust Wide Baseline Stereo From Maximally Stable Extremal Regions.

Siméoni, Avrithis and Chum. CVPR 2019. Local Features and Visual Words Emerge in Activations.

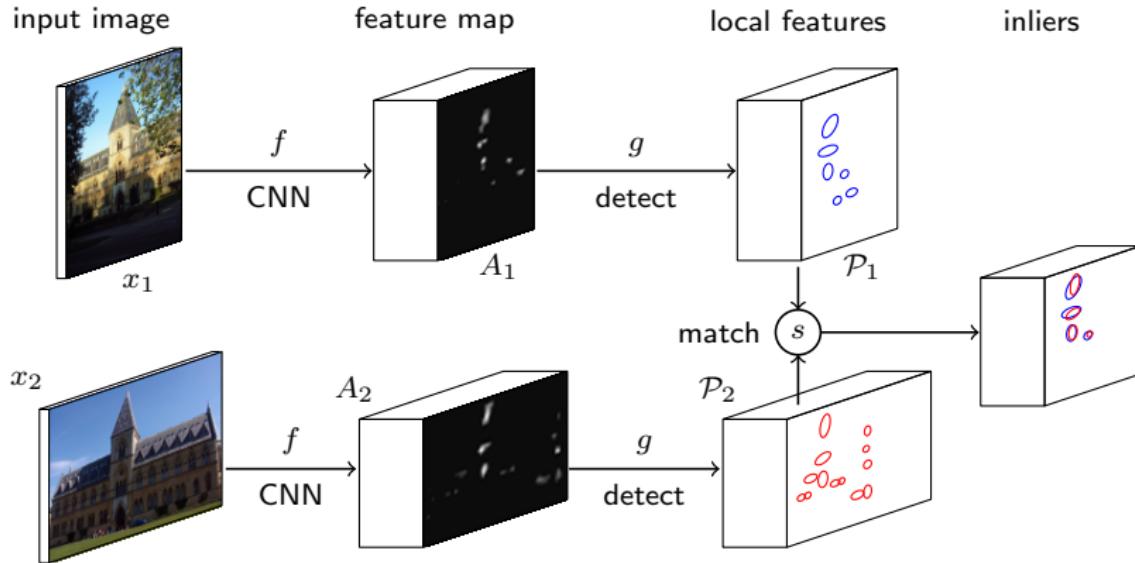
# deep spatial matching (DSM)



- local features detected by MSER independently per channel
- inliers found by fast spatial matching

Philbin, Chum, Isard, Sivic and Zisserman. CVPR 2007. Object Retrieval With Large Vocabularies and Fast Spatial Matching.  
Matas, Chum, Urban and Pajdla. BMVC 2002. Robust Wide Baseline Stereo From Maximally Stable Extremal Regions.  
Siméoni, Avrithis and Chum. CVPR 2019. Local Features and Visual Words Emerge in Activations.

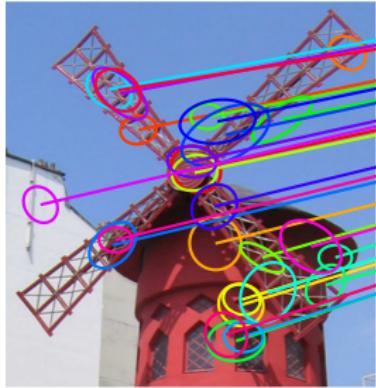
# deep spatial matching (DSM)



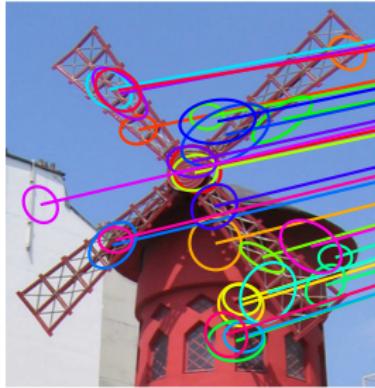
- local features detected by MSER independently per channel
- inliers found by fast spatial matching

Philbin, Chum, Isard, Sivic and Zisserman. CVPR 2007. Object Retrieval With Large Vocabularies and Fast Spatial Matching.  
Matas, Chum, Urban and Pajdla. BMVC 2002. Robust Wide Baseline Stereo From Maximally Stable Extremal Regions.  
Siméoni, Avrithis and Chum. CVPR 2019. Local Features and Visual Words Emerge in Activations.

## example



## example



- local maxima on each activation channel are “local features”
  - channels are “visual words” - no vocabulary needed

# results

## mAP on RevOP using diffusion

Method	Medium		Hard	
	$\mathcal{R}Oxf$	$+\mathcal{R}1M$	$\mathcal{R}Par$	$+\mathcal{R}1M$
V-MAC*	67.7	56.8	39.8	29.4
V-MAC*+DSM	72.0	59.2	43.9	32.0
R-MAC*↑	73.9	61.3	45.6	31.9
R-MAC*↑+DSM	76.9	65.7	49.4	35.7
V-GeM	69.6	60.4	41.1	33.1
V-GeM+DSM	72.8	63.2	45.4	35.4
R-GeM↑	70.1	67.5	41.5	39.6
R-GeM↑+DSM	75.0	70.2	46.2	41.9

- V: VGG-16, R: ResNet-101
- MAC: max-pooling, GeM: generalized mean pooling

# outline – part II

7 context

8 searching on manifolds

9 spatial matching

10 discovering objects

# from attention to detection



## object proposals

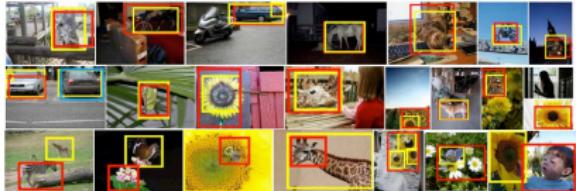
- class-agnostic objectness measure
- essential component of modern two-stage object detectors

# from attention to detection



## object proposals

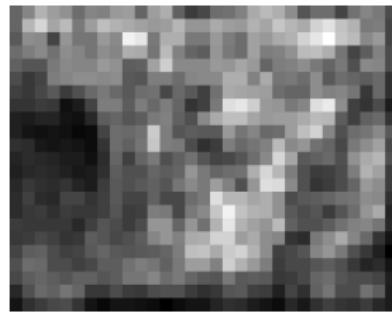
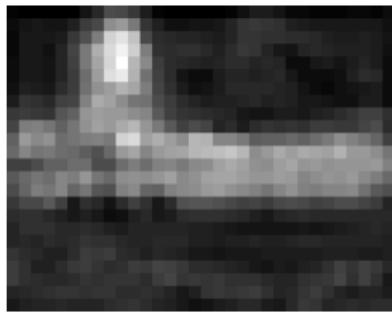
- class-agnostic objectness measure
- essential component of modern two-stage object detectors



## unsupervised object discovery

- segmentation-based ROIs
- rank by link analysis on entire dataset (PageRank)

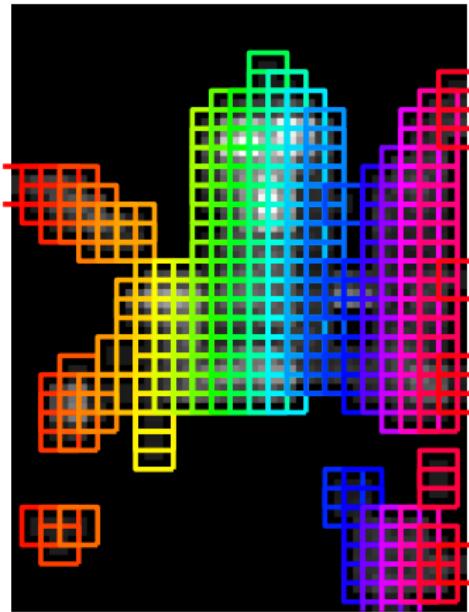
# feature saliency (FS) map



- sparsity-sensitive channel weights on convolutional activations

Kalantidis, Mellina, Osindero. ECCVW 2016. Cross-Dimensional Weighting for Aggregated Deep Convolutional Features.  
Siméoni, Iscen, Tolias, Avrithis, Chum. WACV 2018. Unsupervised deep object discovery for instance recognition.

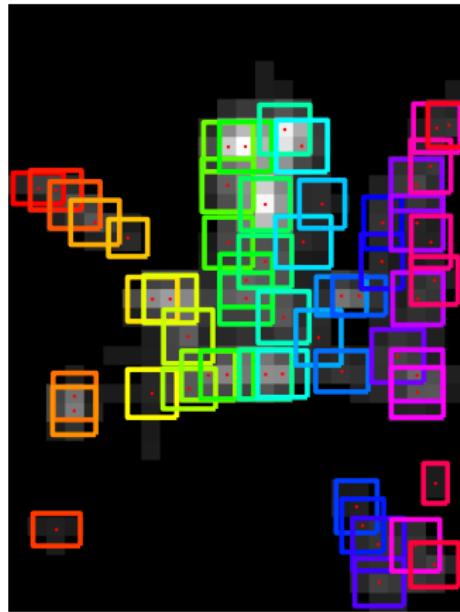
# region detection with EGM



- EGM generalized from points to 2d functions (images)

Avrithis and Kalantidis. ECCV 2012. Approximate Gaussian Mixtures for Large Scale Vocabularies.  
Siméoni, Iscen, Tolias, Avrithis, Chum. WACV 2018. Unsupervised deep object discovery for instance recognition.

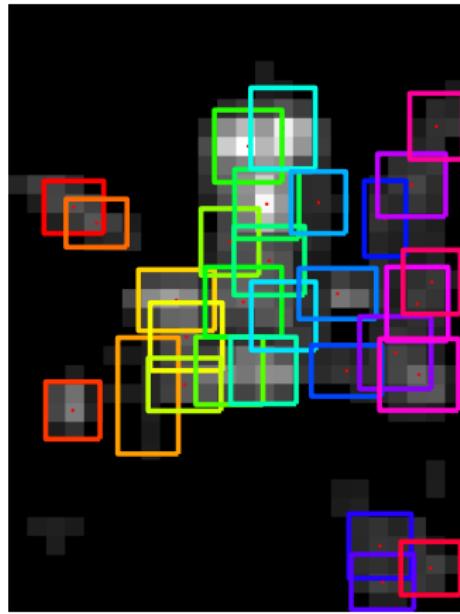
# region detection with EGM



- EGM generalized from points to 2d functions (images)

Avrithis and Kalantidis. ECCV 2012. Approximate Gaussian Mixtures for Large Scale Vocabularies.  
Siméoni, Iscen, Tolias, Avrithis, Chum. WACV 2018. Unsupervised deep object discovery for instance recognition.

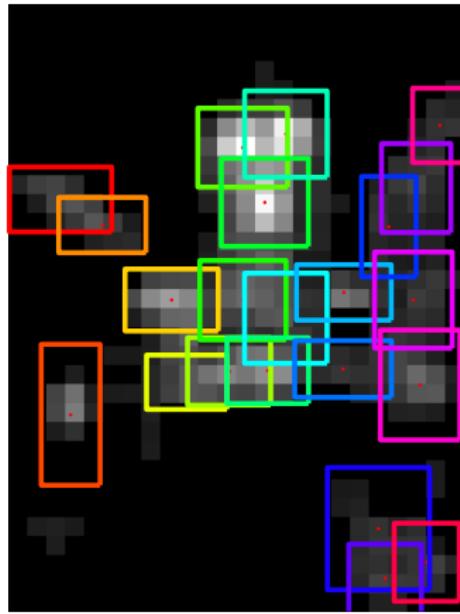
# region detection with EGM



- EGM generalized from points to 2d functions (images)

Avrithis and Kalantidis. ECCV 2012. Approximate Gaussian Mixtures for Large Scale Vocabularies.  
Siméoni, Iscen, Tolias, Avrithis, Chum. WACV 2018. Unsupervised deep object discovery for instance recognition.

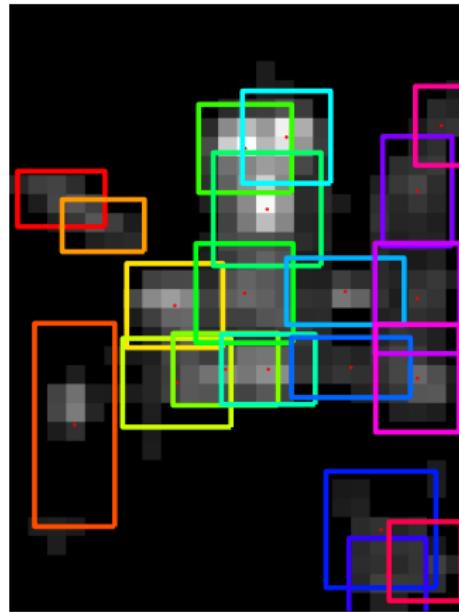
# region detection with EGM



- EGM generalized from points to 2d functions (images)

Avrithis and Kalantidis. ECCV 2012. Approximate Gaussian Mixtures for Large Scale Vocabularies.  
Siméoni, Iscen, Tolias, Avrithis, Chum. WACV 2018. Unsupervised deep object discovery for instance recognition.

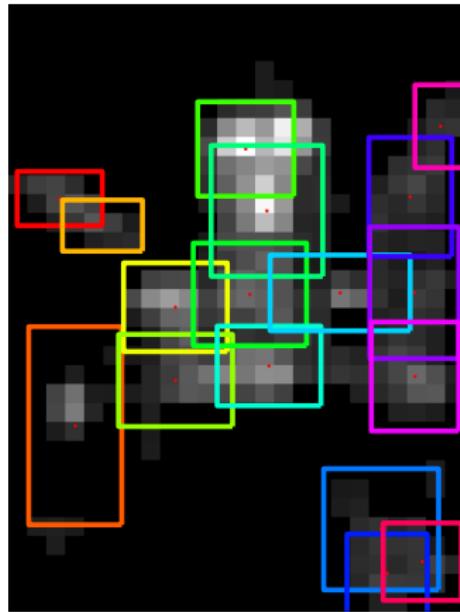
# region detection with EGM



- EGM generalized from points to 2d functions (images)

Avrithis and Kalantidis. ECCV 2012. Approximate Gaussian Mixtures for Large Scale Vocabularies.  
Siméoni, Iscen, Tolias, Avrithis, Chum. WACV 2018. Unsupervised deep object discovery for instance recognition.

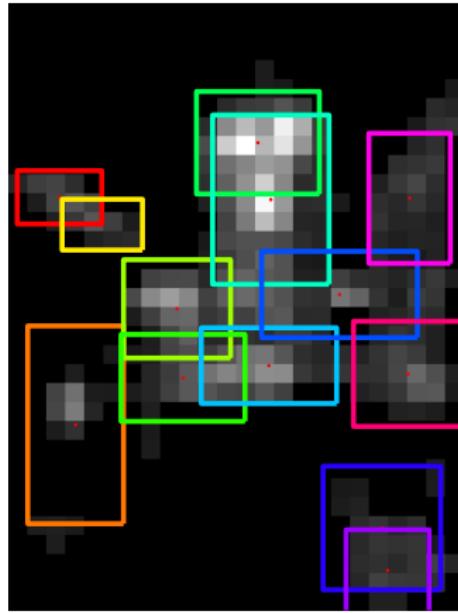
# region detection with EGM



- EGM generalized from points to 2d functions (images)

Avrithis and Kalantidis. ECCV 2012. Approximate Gaussian Mixtures for Large Scale Vocabularies.  
Siméoni, Iscen, Tolias, Avrithis, Chum. WACV 2018. Unsupervised deep object discovery for instance recognition.

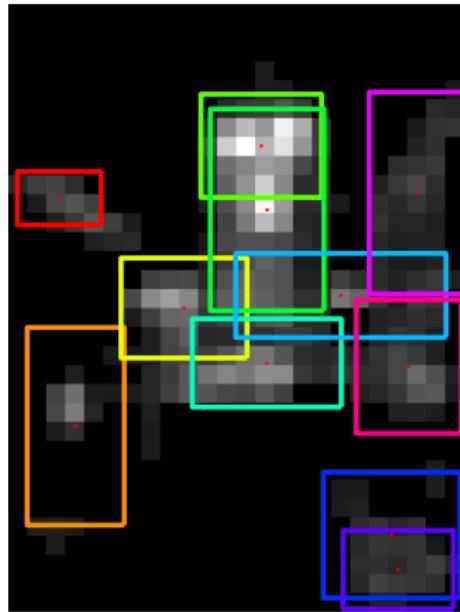
# region detection with EGM



- EGM generalized from points to 2d functions (images)

Avrithis and Kalantidis. ECCV 2012. Approximate Gaussian Mixtures for Large Scale Vocabularies.  
Siméoni, Iscen, Tolias, Avrithis, Chum. WACV 2018. Unsupervised deep object discovery for instance recognition.

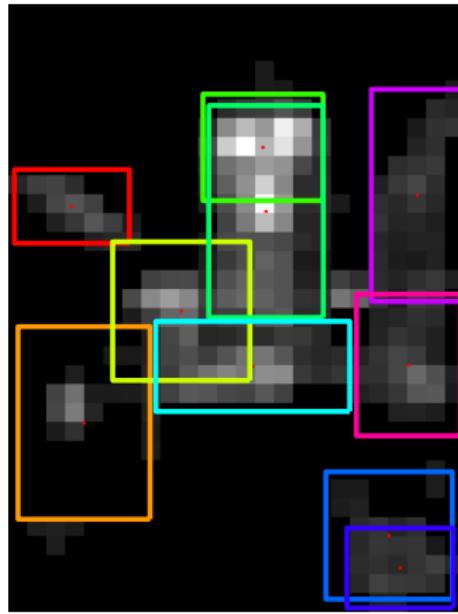
# region detection with EGM



- EGM generalized from points to 2d functions (images)

Avrithis and Kalantidis. ECCV 2012. Approximate Gaussian Mixtures for Large Scale Vocabularies.  
Siméoni, Iscen, Tolias, Avrithis, Chum. WACV 2018. Unsupervised deep object discovery for instance recognition.

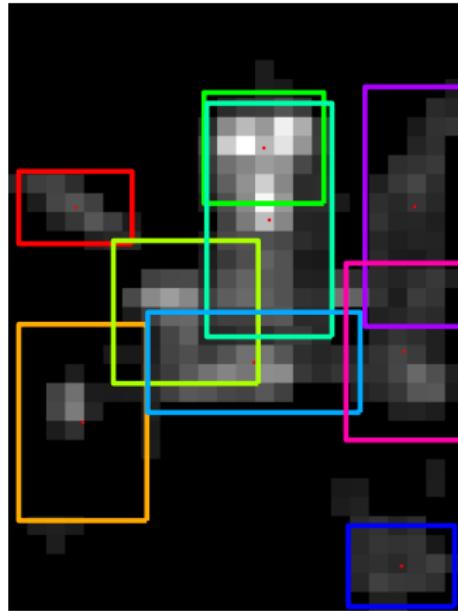
# region detection with EGM



- EGM generalized from points to 2d functions (images)

Avrithis and Kalantidis. ECCV 2012. Approximate Gaussian Mixtures for Large Scale Vocabularies.  
Siméoni, Iscen, Tolias, Avrithis, Chum. WACV 2018. Unsupervised deep object discovery for instance recognition.

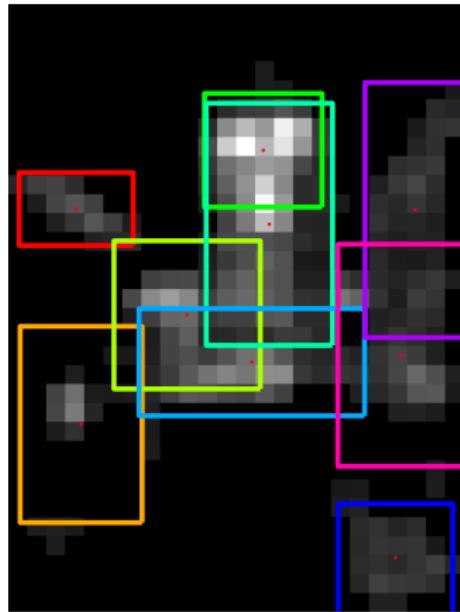
# region detection with EGM



- EGM generalized from points to 2d functions (images)

Avrithis and Kalantidis. ECCV 2012. Approximate Gaussian Mixtures for Large Scale Vocabularies.  
Siméoni, Iscen, Tolias, Avrithis, Chum. WACV 2018. Unsupervised deep object discovery for instance recognition.

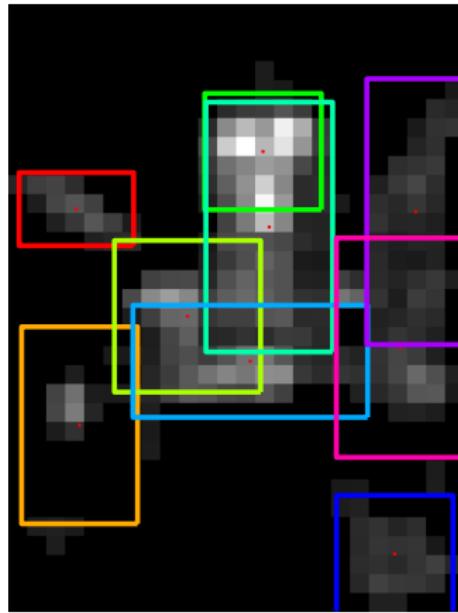
# region detection with EGM



- EGM generalized from points to 2d functions (images)

Avrithis and Kalantidis. ECCV 2012. Approximate Gaussian Mixtures for Large Scale Vocabularies.  
Siméoni, Iscen, Tolias, Avrithis, Chum. WACV 2018. Unsupervised deep object discovery for instance recognition.

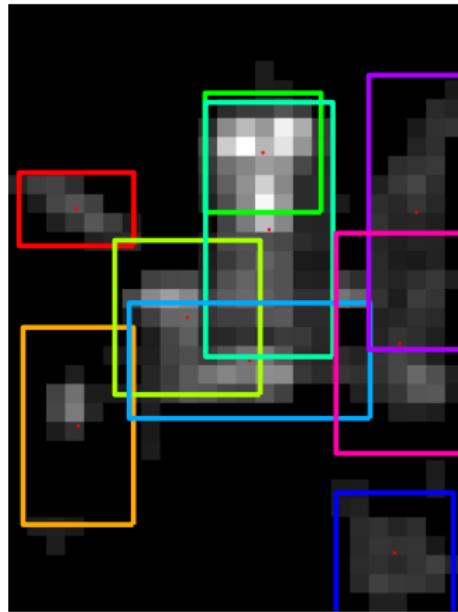
# region detection with EGM



- EGM generalized from points to 2d functions (images)

Avrithis and Kalantidis. ECCV 2012. Approximate Gaussian Mixtures for Large Scale Vocabularies.  
Siméoni, Iscen, Tolias, Avrithis, Chum. WACV 2018. Unsupervised deep object discovery for instance recognition.

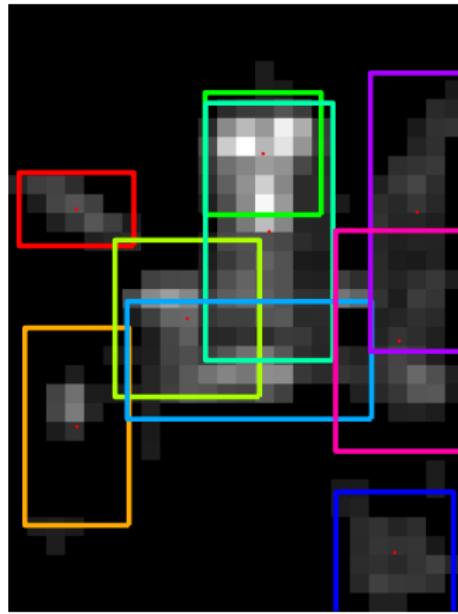
# region detection with EGM



- EGM generalized from points to 2d functions (images)

Avrithis and Kalantidis. ECCV 2012. Approximate Gaussian Mixtures for Large Scale Vocabularies.  
Siméoni, Iscen, Tolias, Avrithis, Chum. WACV 2018. Unsupervised deep object discovery for instance recognition.

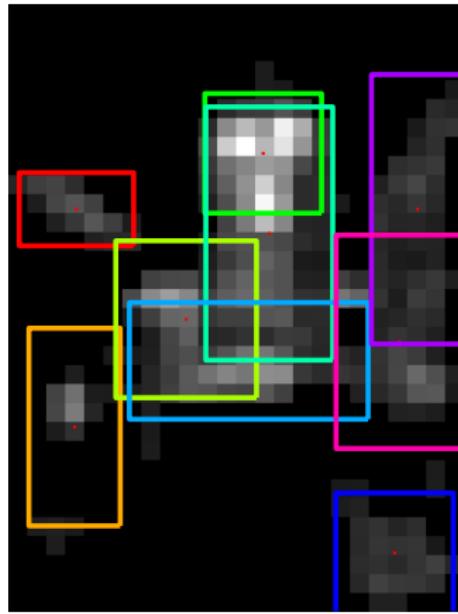
# region detection with EGM



- EGM generalized from points to 2d functions (images)

Avrithis and Kalantidis. ECCV 2012. Approximate Gaussian Mixtures for Large Scale Vocabularies.  
Siméoni, Iscen, Tolias, Avrithis, Chum. WACV 2018. Unsupervised deep object discovery for instance recognition.

# region detection with EGM



- EGM generalized from points to 2d functions (images)

Avrithis and Kalantidis. ECCV 2012. Approximate Gaussian Mixtures for Large Scale Vocabularies.  
Siméoni, Iscen, Tolias, Avrithis, Chum. WACV 2018. Unsupervised deep object discovery for instance recognition.

# object saliency (OS) map

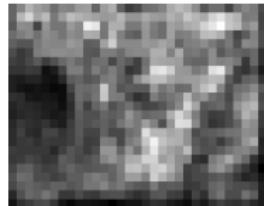
image



- centrality extended to unseen image patches by non-parametric regression

# object saliency (OS) map

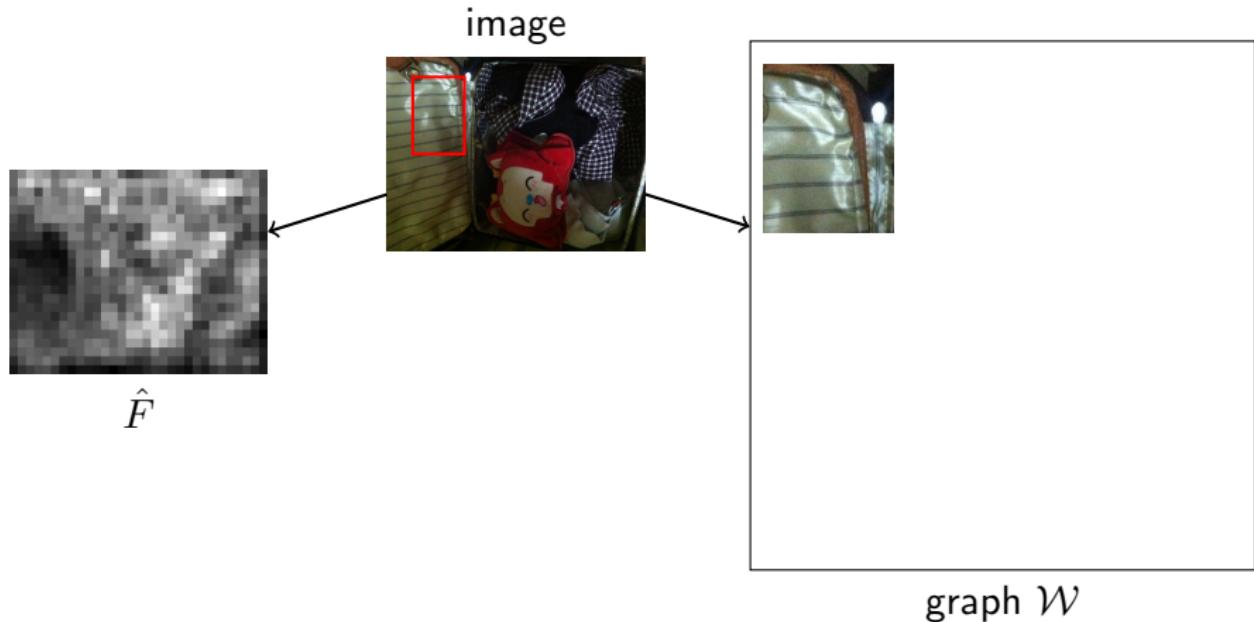
image



$\hat{F}$

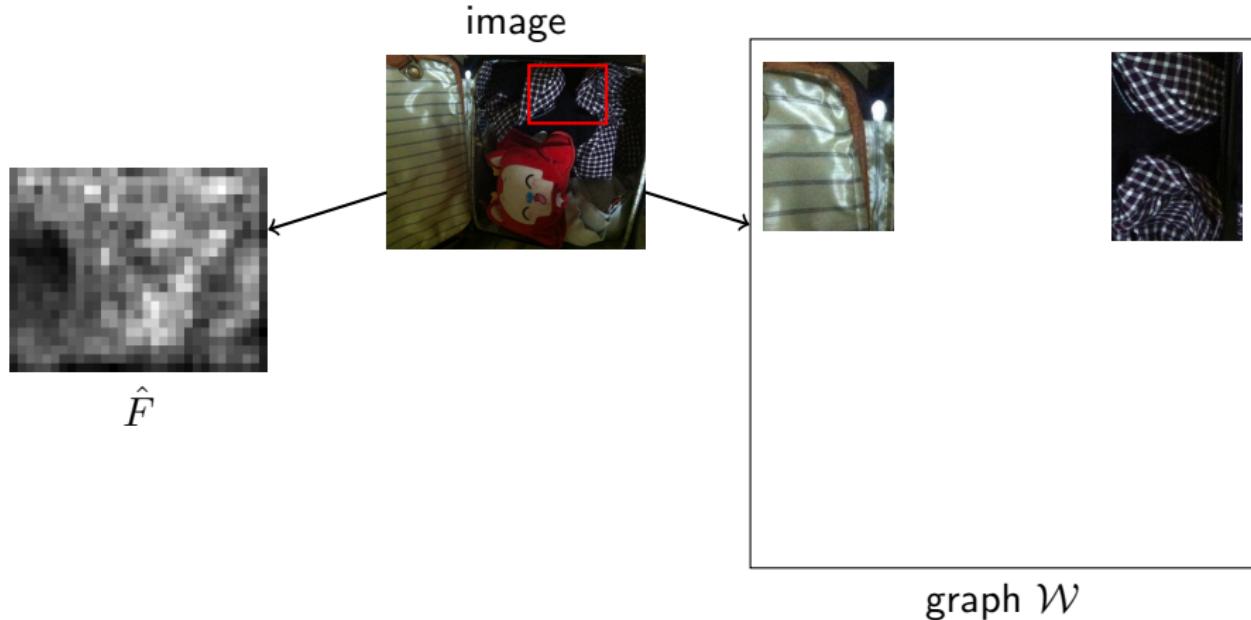
- centrality extended to unseen image patches by non-parametric regression

# object saliency (OS) map



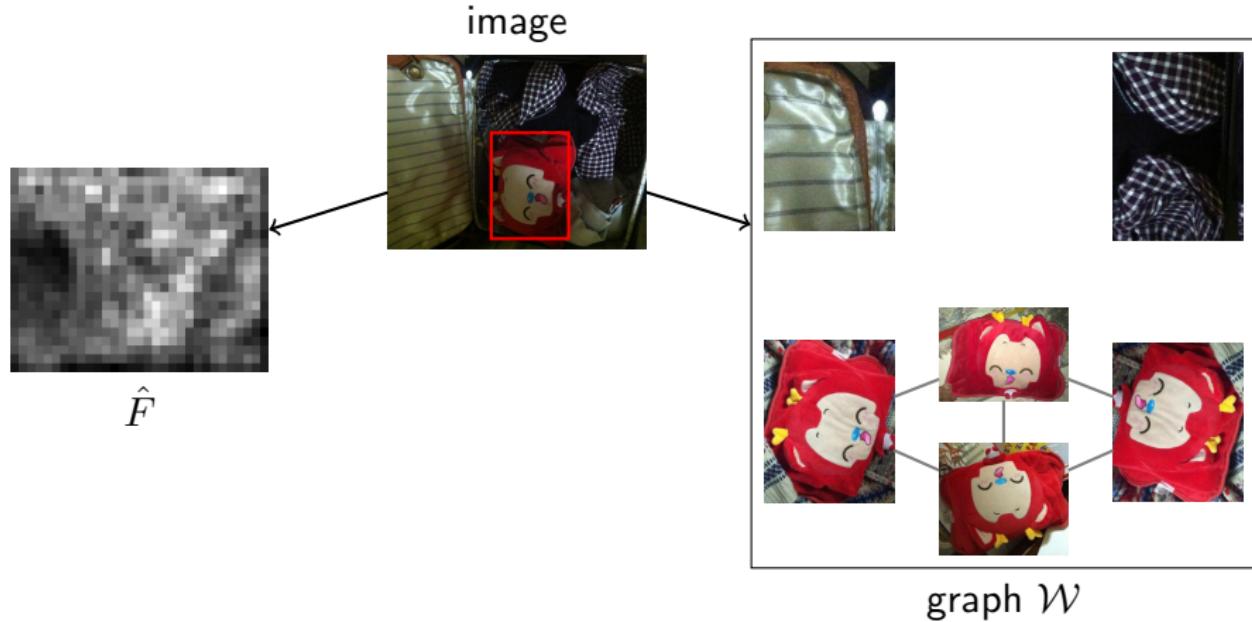
- centrality extended to unseen image patches by non-parametric regression

# object saliency (OS) map



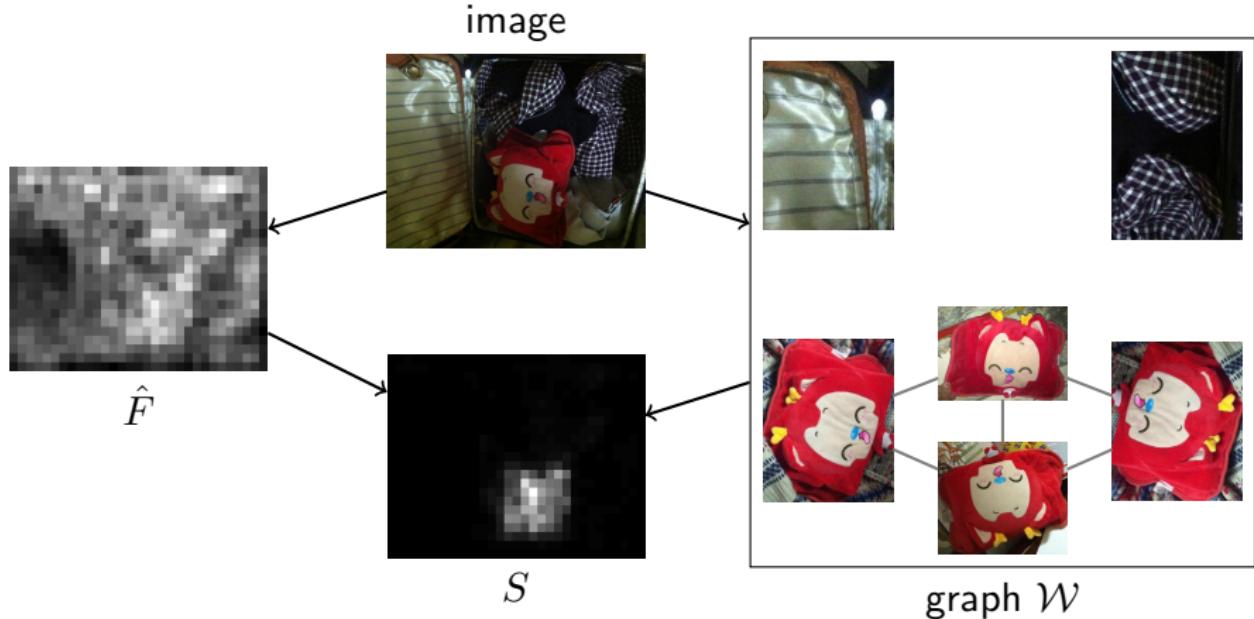
- centrality extended to unseen image patches by non-parametric regression

# object saliency (OS) map



- centrality extended to unseen image patches by non-parametric regression

## object saliency (OS) map



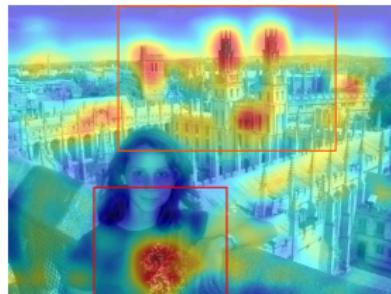
- centrality extended to unseen image patches by non-parametric regression

# FS vs. OS

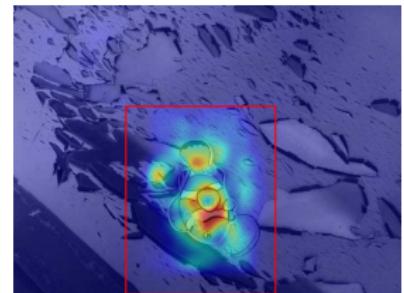
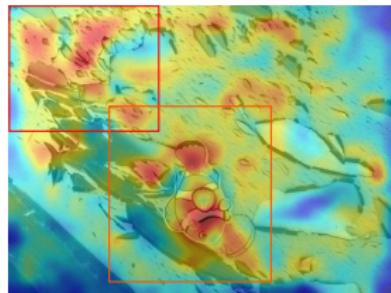
image



FS



os



# results

## mAP on Instre and RevOP using global features

Method	Medium		Hard		
	Instre	$\mathcal{R}\text{Oxf}$	$\mathcal{R}\text{Par}$	$\mathcal{R}\text{Oxf}$	$\mathcal{R}\text{Par}$
GeM	57.0	62.0	69.3	33.7	44.3
FS.EGM	57.7	63.0	68.7	34.5	43.9
OS.EGM	<b>61.3</b>	<b>64.2</b>	<b>69.9</b>	<b>35.9</b>	<b>46.1</b>

- **global** features, pooled from FS/OS regions
- helps particularly on Instre, which contains small objects on background clutter

## achievements and more challenges

- efficient manifold search
- manifold search as smoothing, space-time trade-off
- new retrieval benchmark
- local features emerge without training or altering the architecture
- consistent global and local representations
- suppressing background clutter, without supervision
- dataset-wide analysis improves image representation
- how to learn from minimal data or supervision?

## achievements and more challenges

- efficient manifold search
- manifold search as smoothing, space-time trade-off
- new retrieval benchmark
- local features emerge without training or altering the architecture
- consistent global and local representations
- suppressing background clutter, without supervision
- dataset-wide analysis improves image representation
- how to learn from minimal **data** or **supervision?**

part III

learning

# outline – part III

11 context

12 metric learning

13 semi-supervised learning

14 few-shot learning

# learning with less supervision

## historically

- common (Neocognitron, BoW, layer-wise pre-training)

## in deep learning

- the norm: lots of data, full supervision
- less data/supervision by:
  - autoencoders, generative models
  - transfer learning, domain adaptation
  - proxy tasks: self-supervision, e.g. video, geometric layout, rotation, instance discrimination
  - incremental, few-shot, semi-supervised, weakly-supervised, noisy labels, active learning

# learning with less supervision

## historically

- common (Neocognitron, BoW, layer-wise pre-training)

## in deep learning

- the norm: lots of data, full supervision
- less data/supervision by:
  - autoencoders, generative models
  - transfer learning, domain adaptation
  - proxy tasks: self-supervision, e.g. video, geometric layout, rotation, instance discrimination
  - incremental, few-shot, semi-supervised, weakly-supervised, noisy labels, active learning

## category-level and instance-level tasks converge

- most elements common, e.g. architectures, loss functions, representation learning
- main difference in data and labels, defining **factors of variation** to which invariances need to be learned, e.g.
  - **category-level**: within-class appearance variation
  - **instance-level**: occlusion, clutter, viewpoint changes

# outline – part III

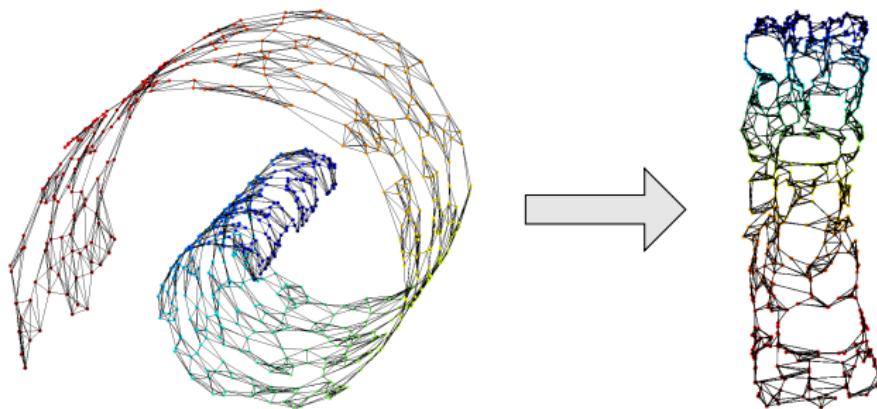
11 context

12 metric learning

13 semi-supervised learning

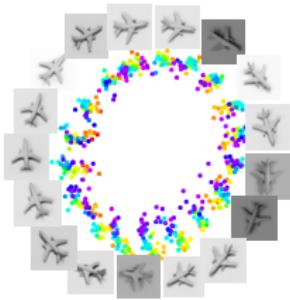
14 few-shot learning

# manifold learning



- classic methods are **unsupervised**
  - do not learn an **explicit mapping** from input to embedding space

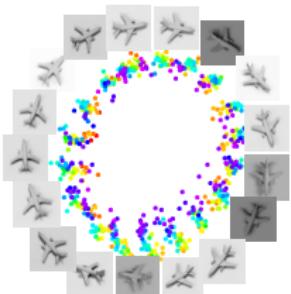
# metric learning



## contrastive learning

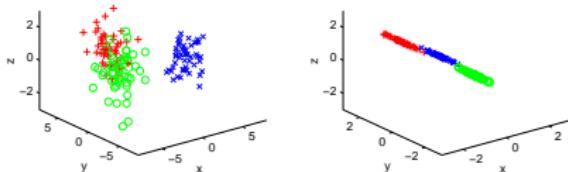
- contrastive loss:  
positive/negative pairs
- unsupervised manifold learning
- explicit nonlinear mapping

# metric learning



## contrastive learning

- contrastive loss:  
positive/negative pairs
- unsupervised manifold learning
- explicit nonlinear mapping



## supervised metric learning

- linear mapping
- positive/negative pairs defined according to class labels

Hadsell, Chopra, Lecun. CVPR 2006. Dimensionality Reduction By Learning an Invariant Mapping.

Xing, Jordan, Russell and N. NIPS 2003. Distance Metric Learning with Application to Clustering with Side-Information.

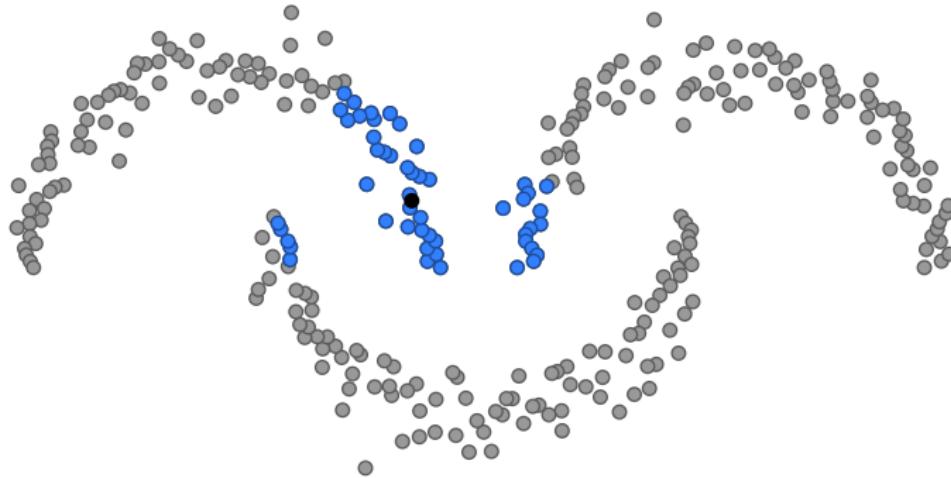
# mining on manifolds (MoM)



- data points ( $\circ$ ), query point  $x$  ( $\bullet$ )

•

## mining on manifolds (MoM)



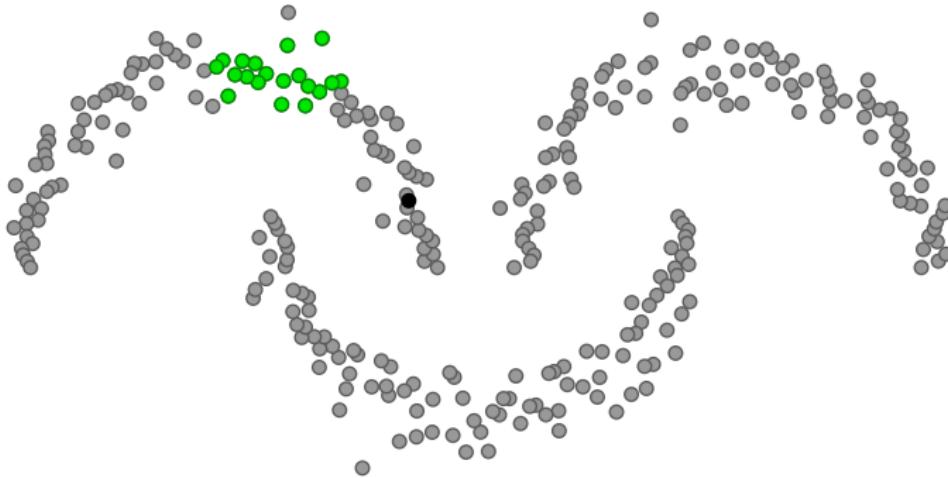
- data points ( $\circ$ ), query point  $x$  ( $\bullet$ )
  - Euclidean nearest neighbors  $E(x)$  ( $\bullet$ )

## mining on manifolds (MoM)



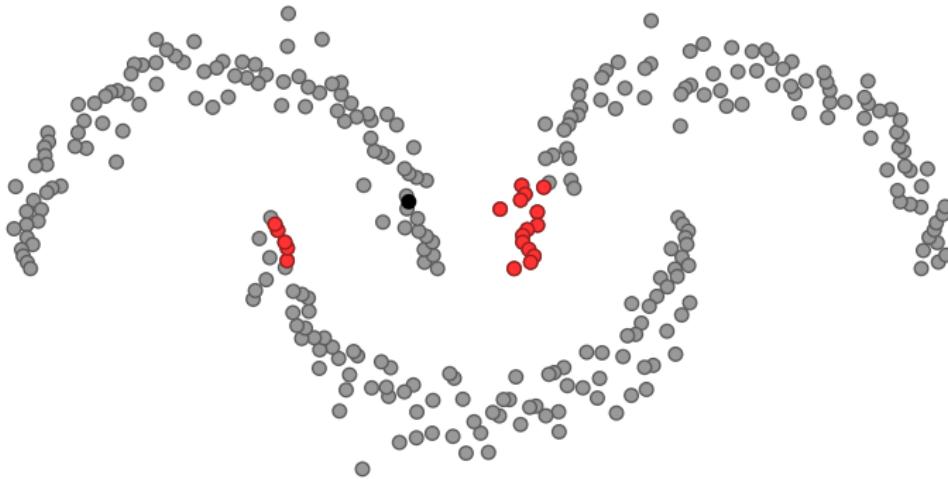
- data points ( $\circ$ ), query point  $x$  ( $\bullet$ )
  - manifold nearest neighbors  $M(x)$  ( $\bullet$ )

# mining on manifolds (MoM)



- data points ( $\circ$ ), query point  $x$  ( $\bullet$ )
- hard positives  $S^+ = M(\mathbf{x}) \setminus E(\mathbf{x})$  ( $\bullet$ )

# mining on manifolds (MoM)



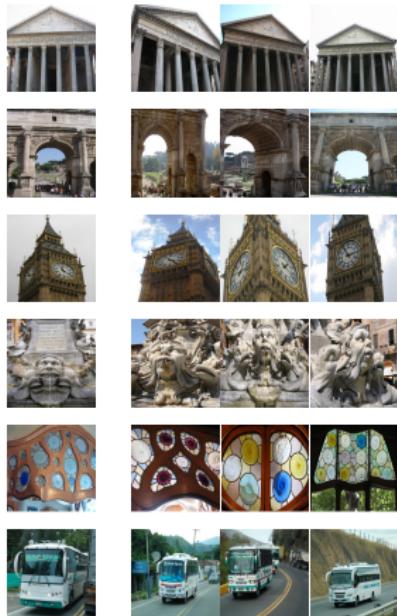
- data points ( $\circ$ ), query point  $x$  ( $\bullet$ )
- hard negatives  $S^- = E(\mathbf{x}) \setminus M(\mathbf{x})$  ( $\bullet$ )

# hard positive/negative examples



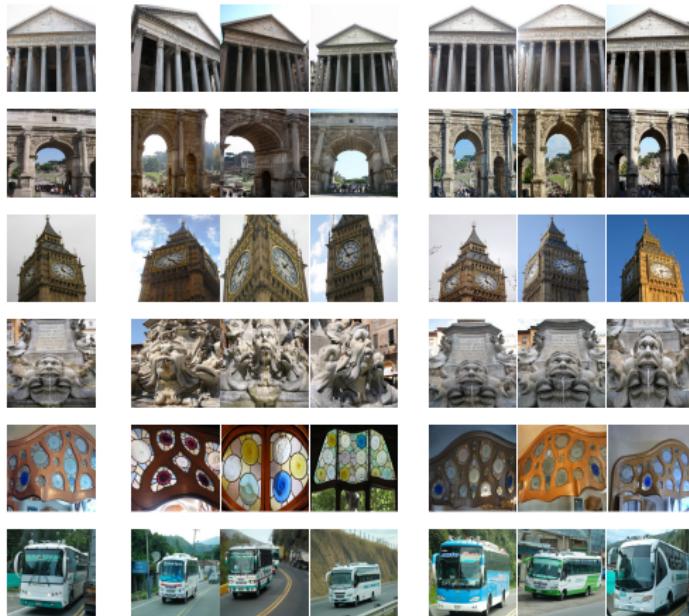
- query (anchor) ( $\mathbf{x}$ )
- positives  $S^+(\mathbf{x})$  vs. Euclidean neighbors  $E(\mathbf{x})$
- negatives  $S^-(\mathbf{x})$  vs. Euclidean non-neighbors  $X \setminus E(\mathbf{x})$

## hard positive/negative examples



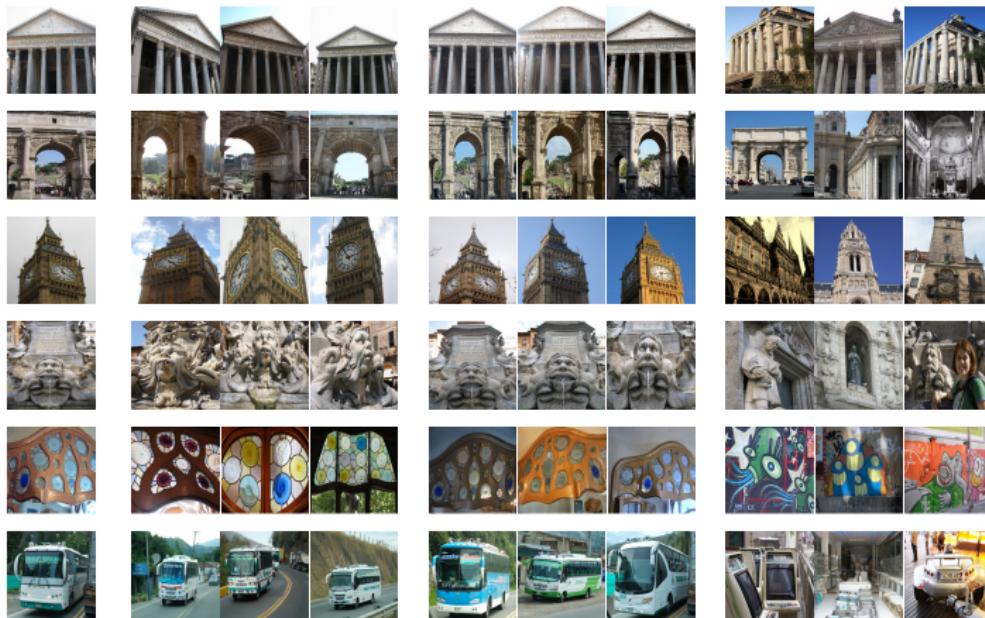
- query (anchor) ( $\mathbf{x}$ )
  - positives  $S^+(\mathbf{x})$  vs. Euclidean neighbors  $E(\mathbf{x})$
  - negatives  $S^-(\mathbf{x})$  vs. Euclidean non-neighbors  $X \setminus E(\mathbf{x})$

## hard positive/negative examples



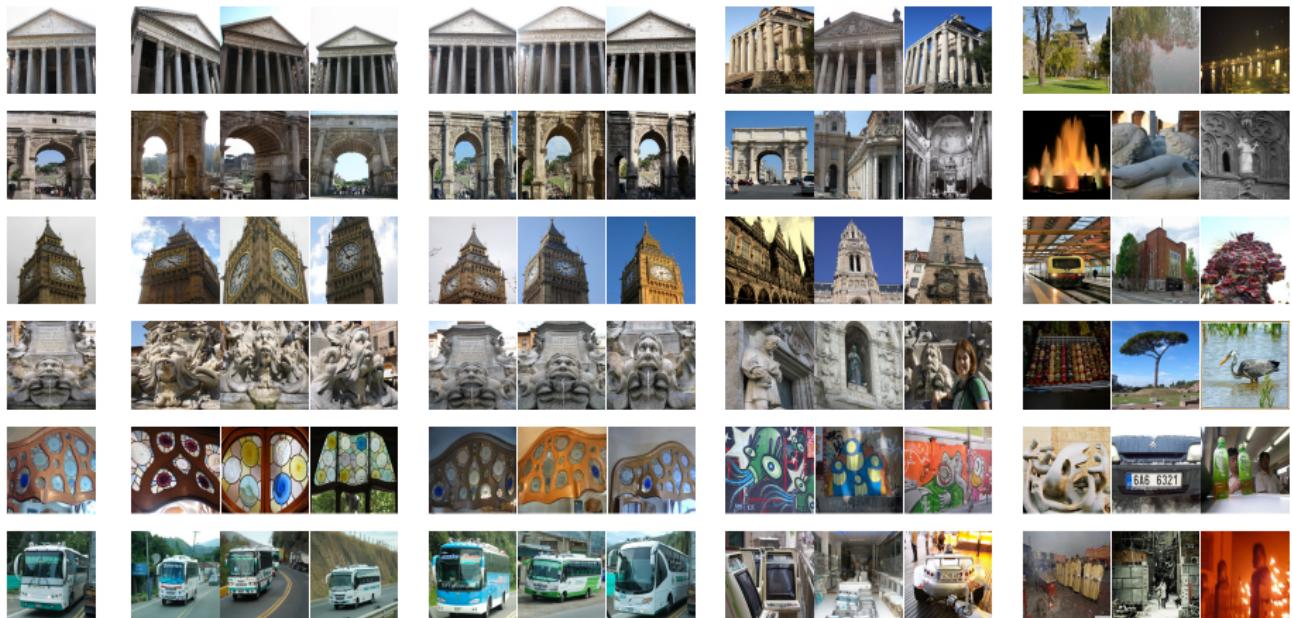
- query (anchor) ( $\mathbf{x}$ )
  - positives  $S^+(\mathbf{x})$  vs. Euclidean neighbors  $E(\mathbf{x})$
  - negatives  $S^-(\mathbf{x})$  vs. Euclidean non-neighbors  $X \setminus E(\mathbf{x})$

## hard positive/negative examples



- query (anchor) ( $\mathbf{x}$ )
  - positives  $S^+(\mathbf{x})$  vs. Euclidean neighbors  $E(\mathbf{x})$
  - negatives  $S^-(\mathbf{x})$  vs. Euclidean non-neighbors  $X \setminus E(\mathbf{x})$

## hard positive/negative examples



- query (anchor) ( $\mathbf{x}$ )
  - positives  $S^+(\mathbf{x})$  vs. Euclidean neighbors  $E(\mathbf{x})$
  - negatives  $S^-(\mathbf{x})$  vs. Euclidean non-neighbors  $X \setminus E(\mathbf{x})$

# results

## fine-grained categorization

Method	Labels	R@1	R@2	R@4	R@8	NMI
Baseline		35.0	46.8	59.3	72.0	48.1
Cyclic match		40.8	52.8	65.1	76.0	52.6
MoM (ours)		<b>45.3</b>	<b>57.8</b>	<b>68.6</b>	<b>78.4</b>	<b>55.0</b>
Triplet+semi-hard	✓	42.3	55.0	66.4	77.2	55.4
Lifted-structure	✓	43.6	56.6	68.6	79.6	56.5
Triplet+	✓	45.9	57.7	69.6	79.8	58.1
Clustering	✓	48.2	61.4	71.8	81.9	59.2
Triplet+++	✓	49.8	62.3	74.1	83.3	59.9

- CUB200-2011 dataset, 200 bird species, 100 training / 100 testing
- GoogLeNet pre-trained on ImageNet, then fine-tuned with triplet loss

# results

## particular object retrieval

Method	Hol	Instre	Oxf5k	Oxf105k	Par6k	Par106k
Testing on MAC						
Baseline	79.4	48.5	58.5	50.3	73.0	59.0
SfM	81.4	48.5	79.7	73.9	82.4	74.6
MoM (ours)	82.6	55.5	78.7	74.3	83.1	75.6
Testing on R-MAC						
Baseline	87.0	55.6	68.0	61.0	76.6	72.1
SfM	84.4	47.7	77.8	70.1	84.1	76.8
MoM (ours)	87.5	57.7	78.2	72.6	85.1	78.0

- VGG-16 pre-trained on ImageNet, then fine-tuned with contrastive loss on a 1M **unlabeled** dataset with MAC pooling

# outline – part III

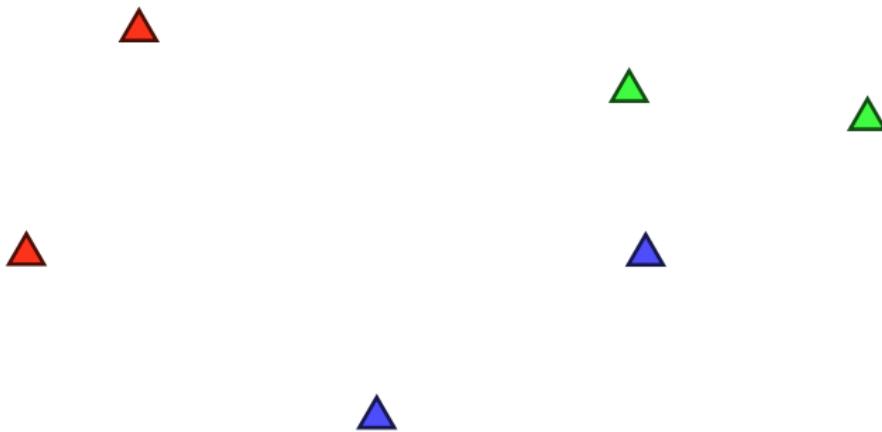
11 context

12 metric learning

13 semi-supervised learning

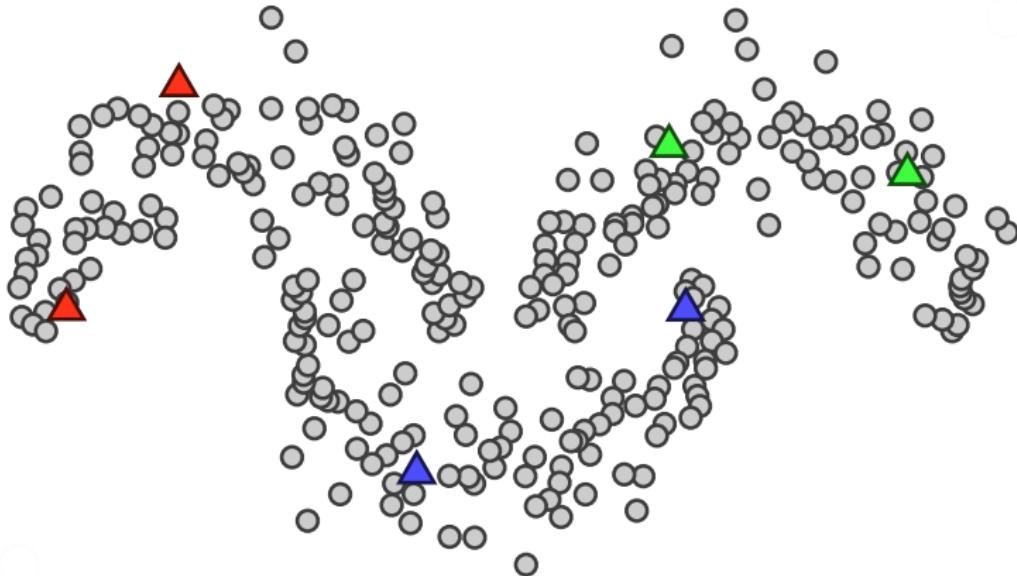
14 few-shot learning

## semi-supervised learning



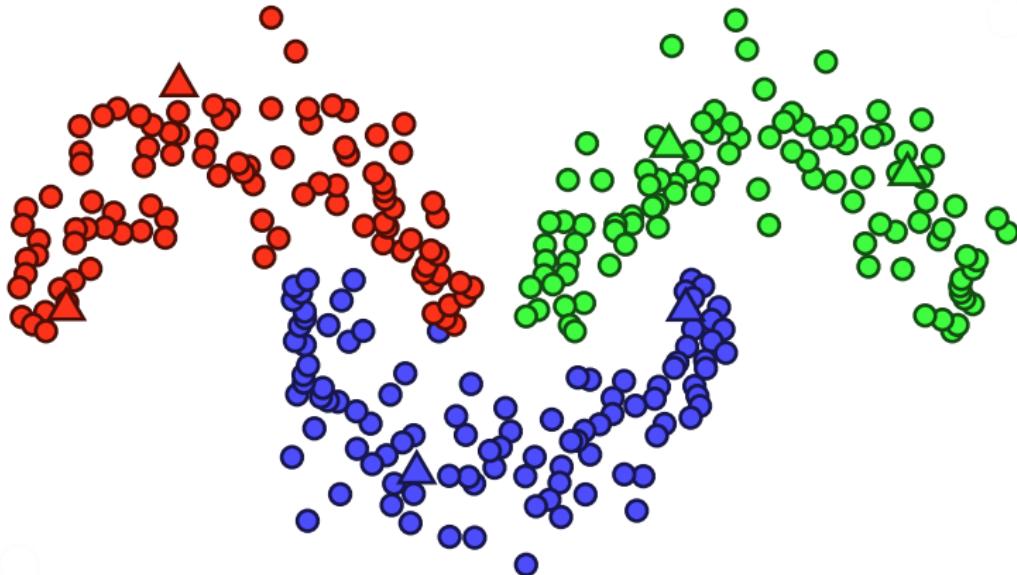
- labeled points ( $\blacktriangle$ ), unlabeled points  $x$  ( $\circ$ )
  - propagated labels ( $\bullet$ ) certainty of prediction

## semi-supervised learning



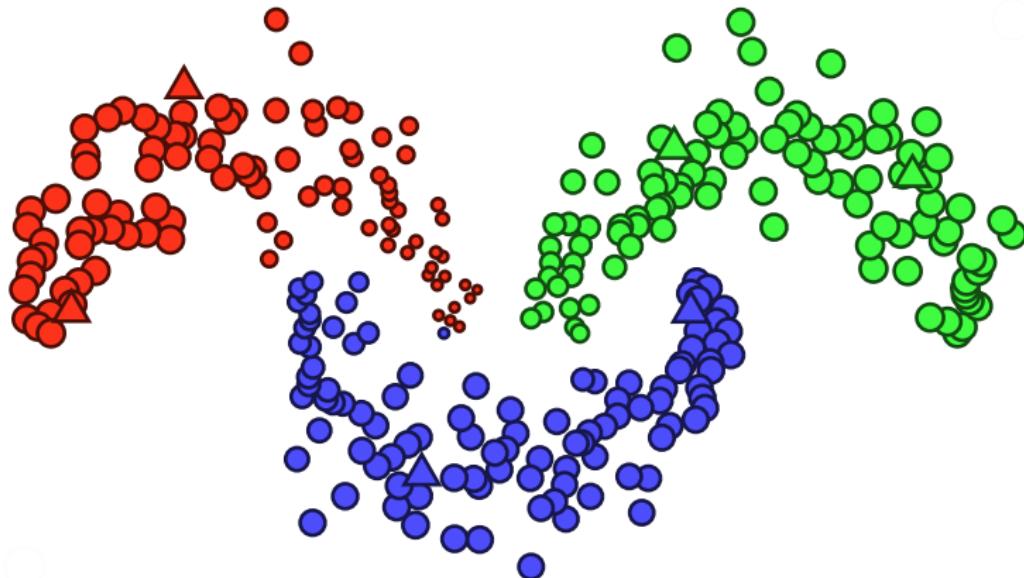
- labeled points ( $\blacktriangle$ ), unlabeled points  $x$  ( $\circ$ )
  - propagated labels ( $\bullet$ ) certainty of prediction

## label propagation (transductive)



- labeled points ( $\blacktriangle$ ), unlabeled points  $x$  ( $\circ$ )
- propagated labels ( $\bullet$ ), certainty of prediction

## label propagation (transductive)



- labeled points ( $\blacktriangle$ ), unlabeled points  $x$  ( $\circ$ )
- propagated labels ( $\bullet$ ), certainty of prediction

# common inductive approaches

$$y'_i = \begin{cases} 1 & \text{if } i = \operatorname{argmax}_{i'} f_{i'}(x) \\ 0 & \text{otherwise} \end{cases}$$

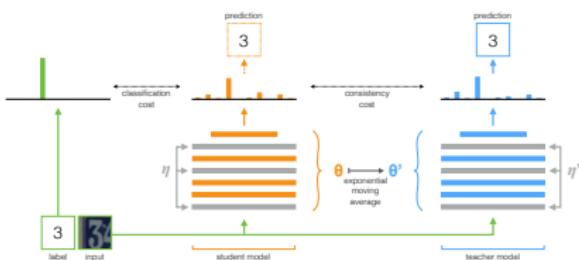
## pseudo-labels

- treat predictions as ground truth
- dates back to the 60's

Lee. WCRL 2013. Pseudo-Label: the Simple and Efficient Semi-Supervised Learning Method for Deep Neural Networks.  
Tsvatinen and Valpola. NIPS 2017. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results.

# common inductive approaches

$$y'_i = \begin{cases} 1 & \text{if } i = \operatorname{argmax}_{i'} f_{i'}(x) \\ 0 & \text{otherwise} \end{cases}$$



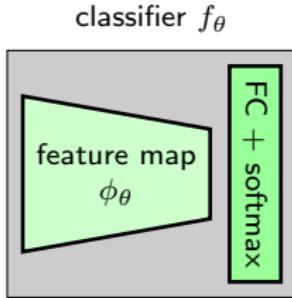
## pseudo-labels

- treat predictions as ground truth
- dates back to the 60's

## consistency losses

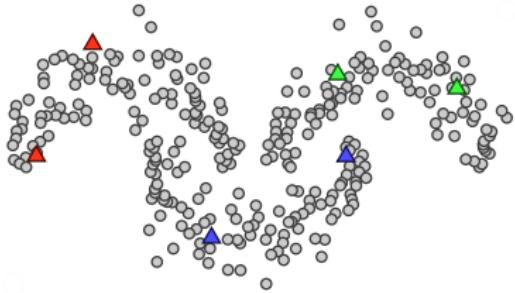
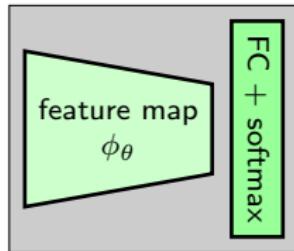
- predictions of similar networks on same input encouraged to be similar

# deep label propagation (DLP) (inductive)

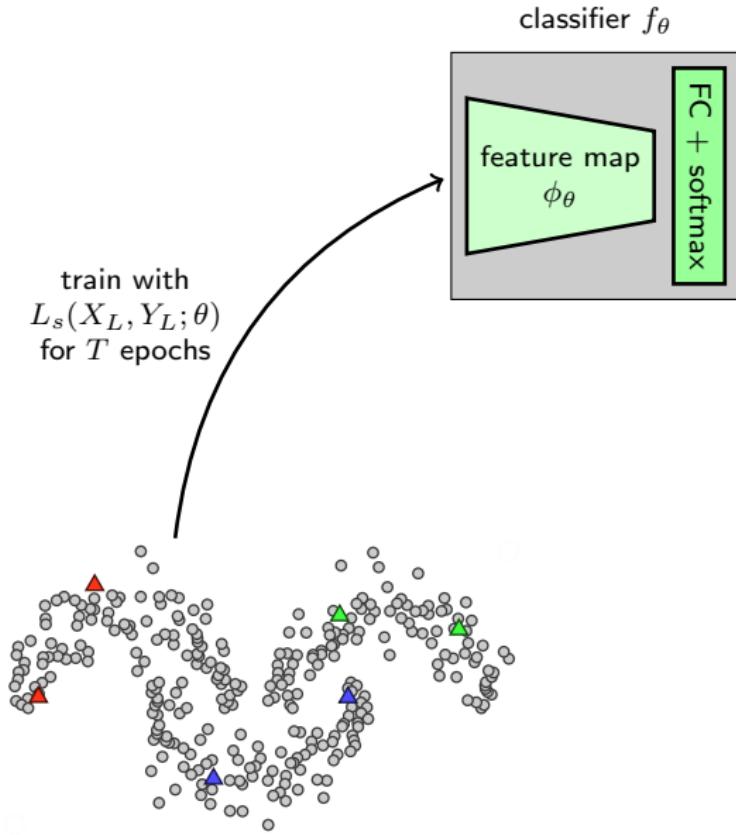


## deep label propagation (DLP) (inductive)

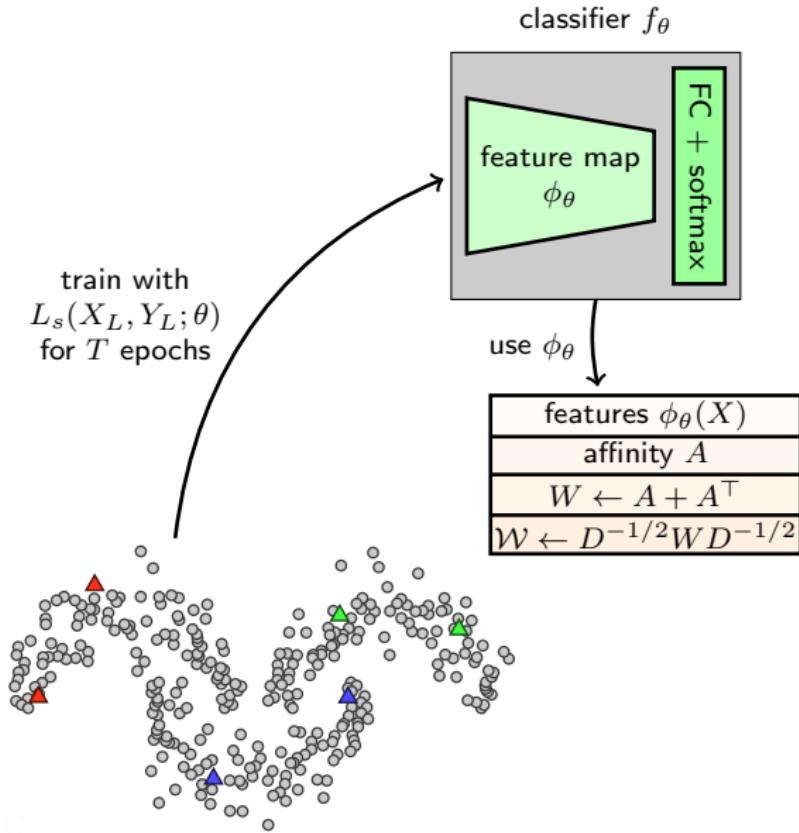
classifier  $f_\theta$



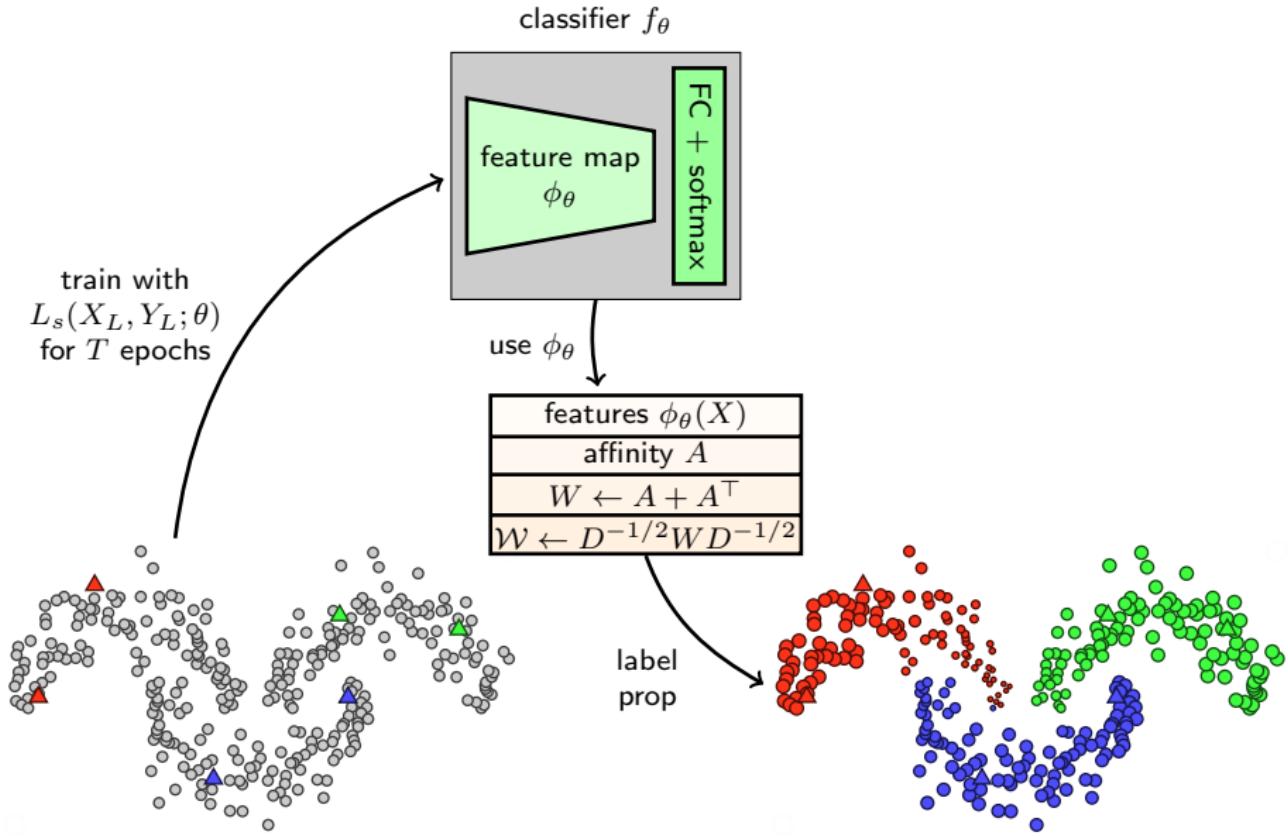
## deep label propagation (DLP) (inductive)



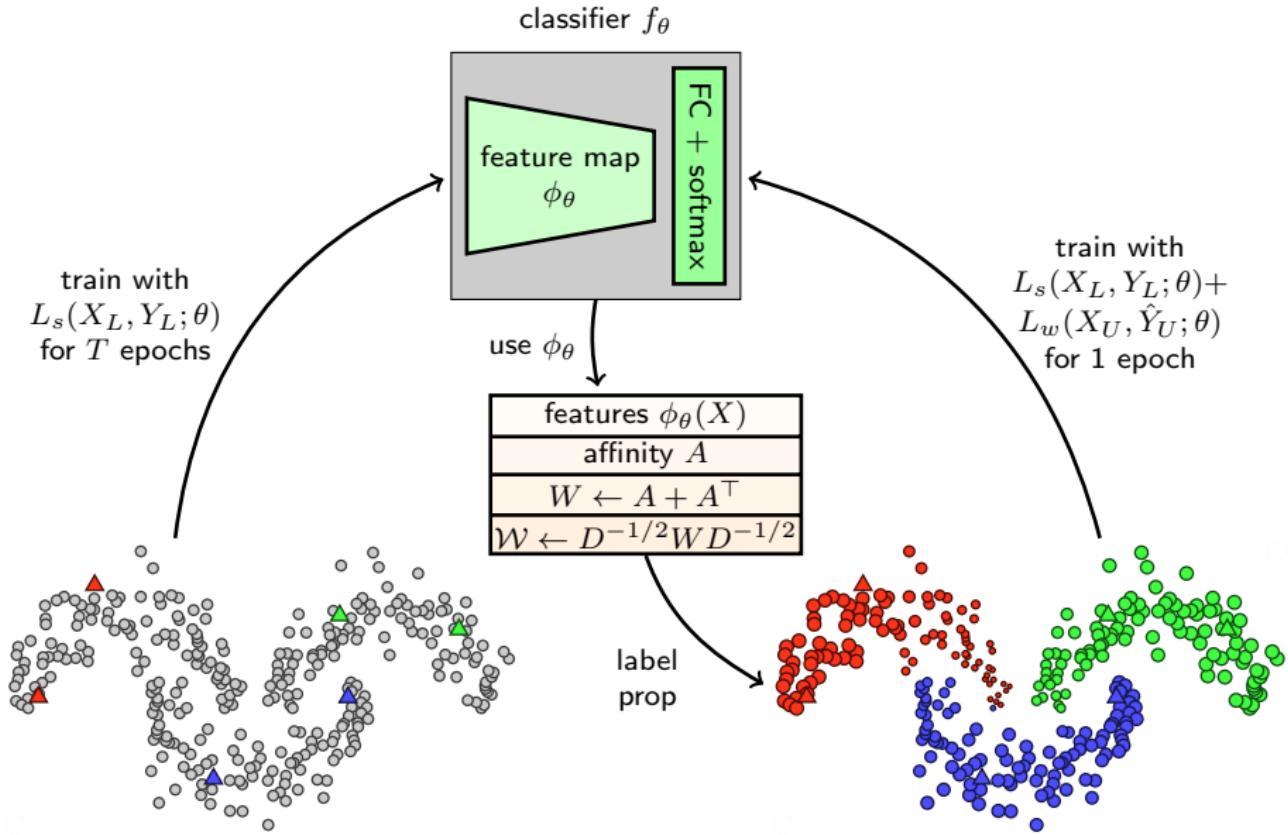
## deep label propagation (DLP) (inductive)



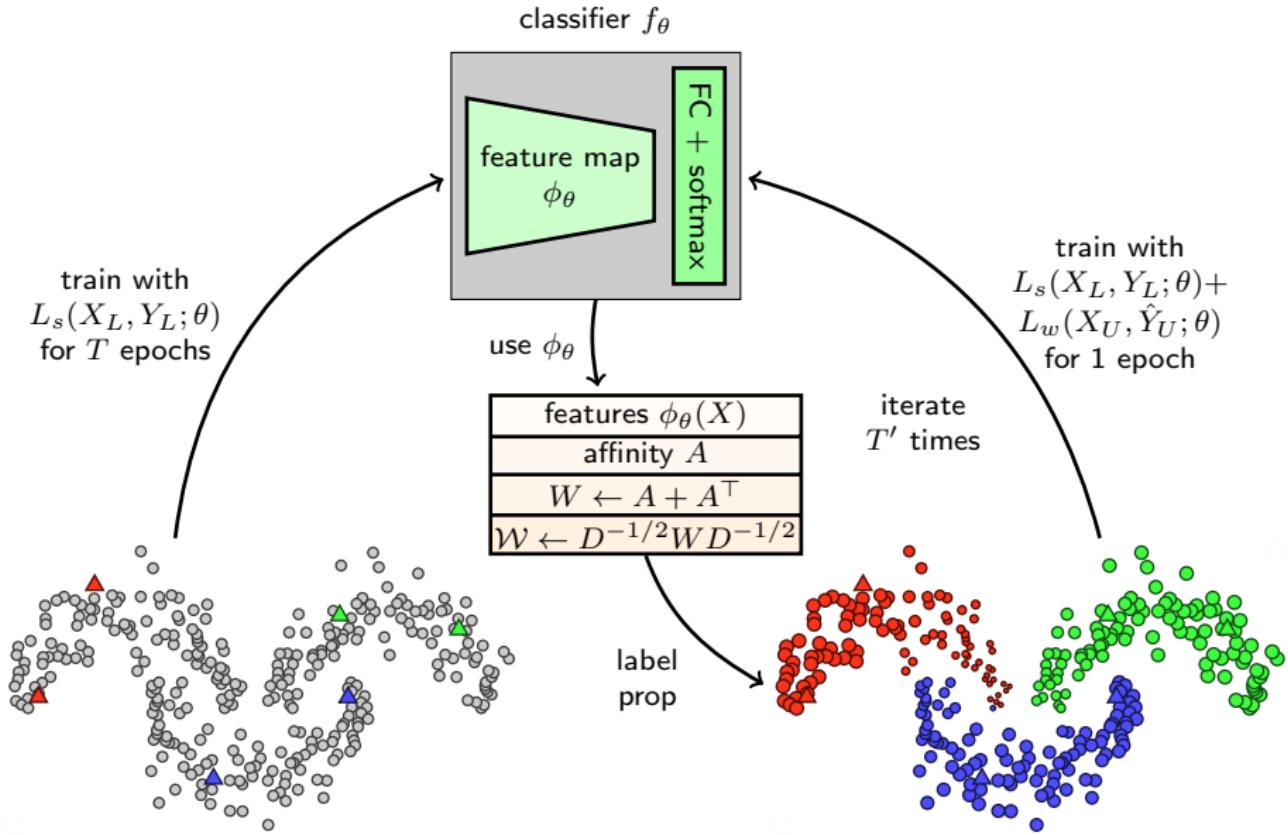
## deep label propagation (DLP) (inductive)



## deep label propagation (DLP) (inductive)



## deep label propagation (DLP) (inductive)



# results

## classification error

Dataset	CIFAR-10		CIFAR-100		<i>minilmageNet</i>	
# Labels	500	1,000	4,000	10,000	4,000	10,000
Supervised	49.08	40.03	55.43	40.67	53.07	38.28
DLP	32.40	22.02	46.20	38.43	47.58	36.14
MT	27.45	19.04	45.36	36.08	49.35	32.51
MT+DLP	24.02	16.93	43.73	35.92	50.52	31.99

- C13 on CIFAR-10/100, ResNet-18 on *minilmageNet*
- either DLP or MT+DLP works best

Tarvainen and Valpola. NIPS 2017. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results.

Iscen, Tolias, Avrithis and Chum. CVPR 2019. Label Propagation for Deep Semi-supervised Learning.

# outline – part III

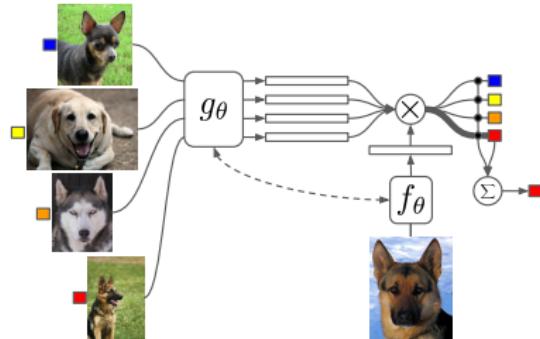
11 context

12 metric learning

13 semi-supervised learning

14 few-shot learning

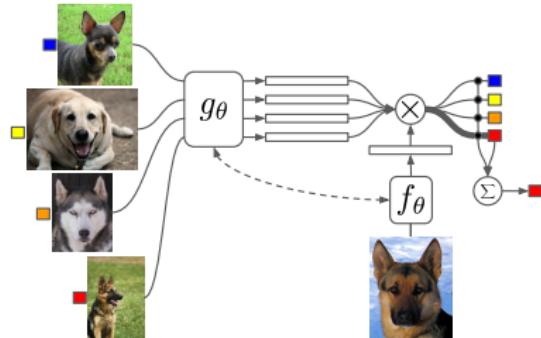
# **few-shot learning**



# metric learning

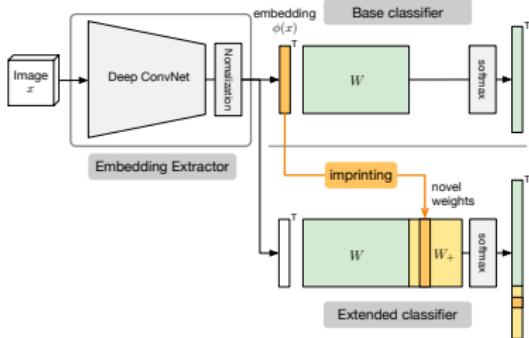
- learn to compare on **base** classes
  - at inference: compare on **novel** classes

# few-shot learning



## metric learning

- learn to compare on **base** classes
- at inference: compare on **novel** classes

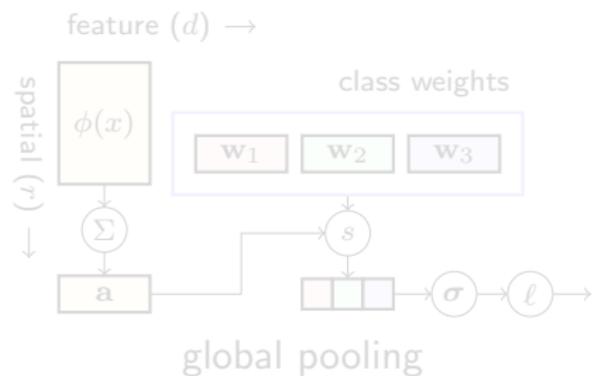
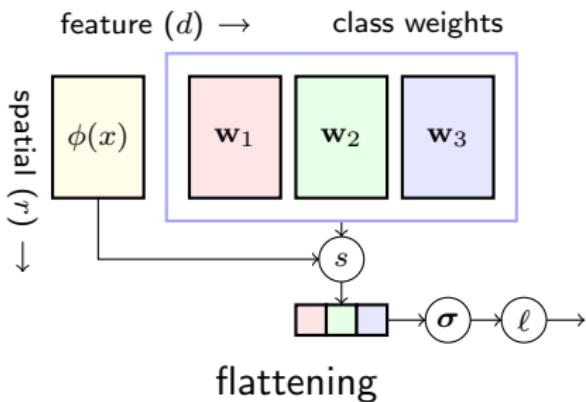


## cosine similarity-based classifier

- features and class weight vectors 2-normalized
- standard **cross-entropy** loss on base classes

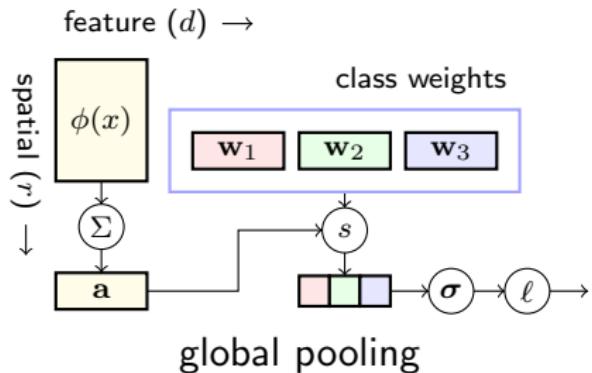
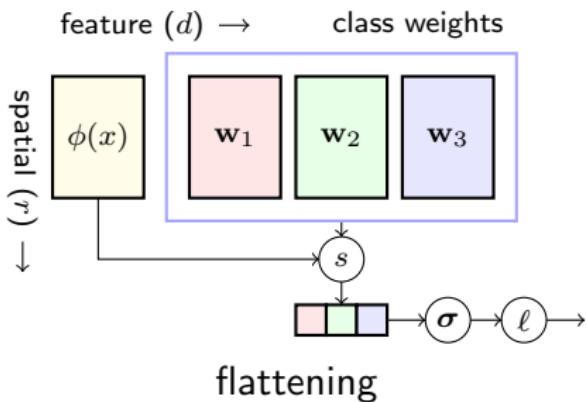
Vinyals, Blundell, Lillicrap, Kavukcuoglu and Wierstra. NIPS 2016. Matching Networks for One-Shot Learning.  
Qi, Brown and Lowe. CVPR 2018. Low-Shot Learning With Imprinted Weights.

# from tensors to vectors



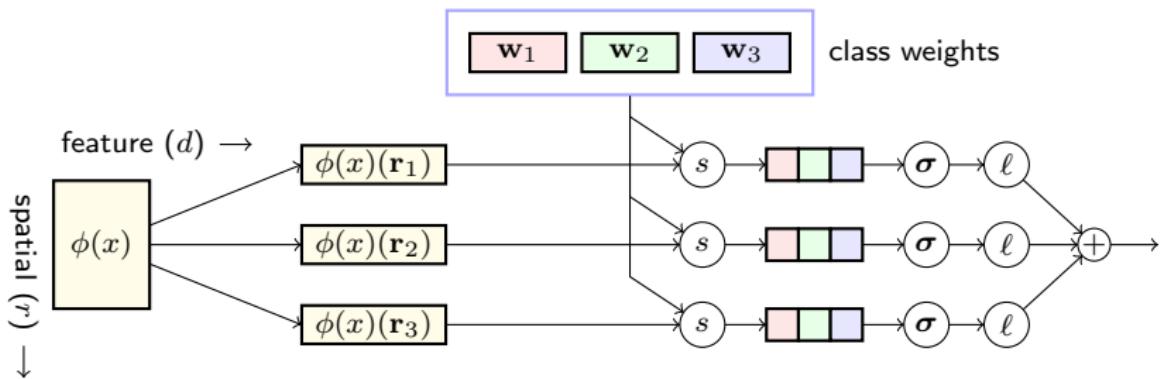
- flattening is very discriminative, but not invariant
- global spatial pooling (GAP) is invariant, but less discriminative

# from tensors to vectors



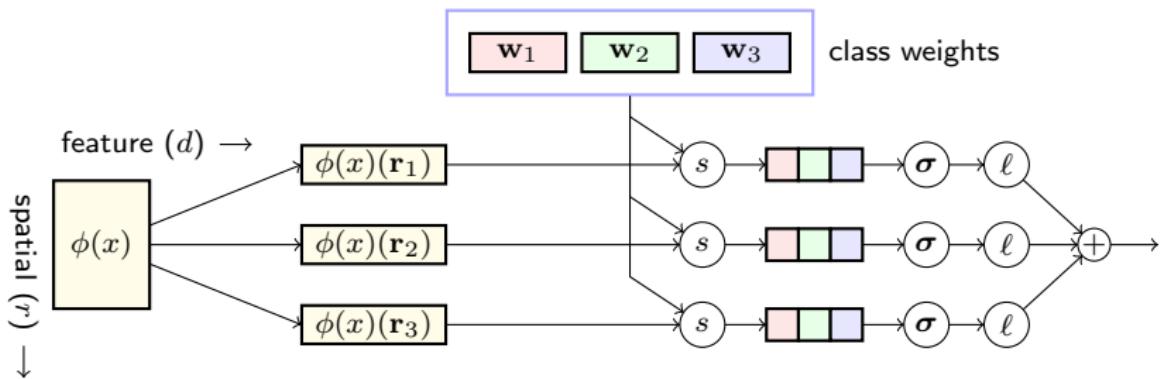
- flattening is very discriminative, but not invariant
- global spatial pooling (GAP) is invariant, but less discriminative

## dense classification (DC)



- $1 \times 1$  convolution followed by depth-wise softmax
  - classifier encouraged to make correct predictions everywhere
  - behaves like implicit data **augmentation** of exhaustive shifts and crops

## dense classification (DC)



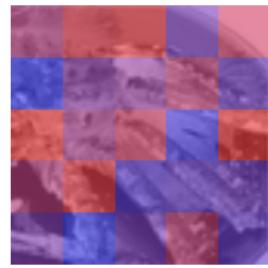
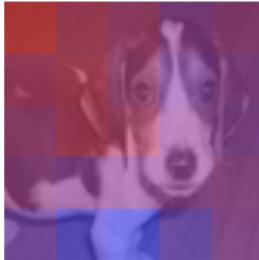
- $1 \times 1$  convolution followed by depth-wise softmax
  - classifier encouraged to make correct predictions everywhere
  - behaves like implicit data **augmentation** of exhaustive shifts and crops

# dense classification (DC)

base classes



novel classes



pooling

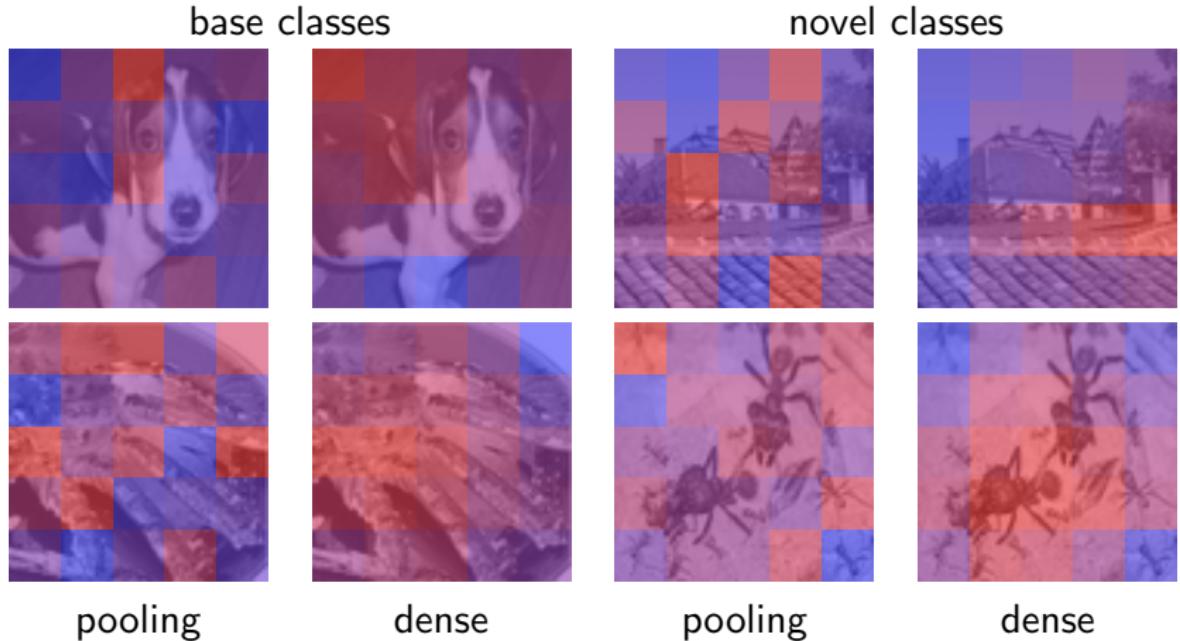
dense

pooling

dense

- blue (red) is low (high) activation for ground truth
- smoother activation maps, more aligned with objects

# dense classification (DC)



- blue (red) is low (high) activation for ground truth
- smoother activation maps, more aligned with objects

# results

## 5-way novel-class classification accuracy on minilmageNet

Method	1-shot	5-shot	10-shot
GAP	$58.61 \pm 0.18$	$76.40 \pm 0.13$	$80.76 \pm 0.11$
DC (ours)	$62.53 \pm 0.19$	$78.95 \pm 0.13$	$82.66 \pm 0.11$
DC + Wide	$61.73 \pm 0.19$	$78.25 \pm 0.14$	$82.03 \pm 0.12$
DC + IMP (ours)	—	$79.77 \pm 0.19$	$83.83 \pm 0.16$
Gidaris <i>et al.</i>	$55.45 \pm 0.70$	$73.00 \pm 0.60$	—
ProtoNet	$56.50 \pm 0.40$	$74.20 \pm 0.20$	$78.60 \pm 0.40$
TADAM	$58.50 \pm 0.30$	$76.70 \pm 0.30$	$80.80 \pm 0.30$

- ResNet-12, following TADAM
- helps particularly on 1-shot

Gidaris and Komodakis. CVPR 2018. Dynamic Few-Shot Visual Learning Without Forgetting.

Oreshkin, Rodriguez, Lacoste. NIPS 2018. TADAM: Task dependent adaptive metric for improved few-shot learning.

Lifchitz, Avrithis, Picard and Bursuc. CVPR 2019. Dense Classification and Implanting for Few-Shot Learning.

# achievements

- revival of unsupervised metric learning
- self-learning without conventional pipelines
- revival of transductive methods and pseudo-labels
- dataset-wide analysis **iteratively** improves image representation
- first study of local activations in few-shot learning
- training to convergence in few-shot learning
- advances on robustness of convolutional networks

# achievements

- revival of unsupervised metric learning
- self-learning without conventional pipelines
- revival of transductive methods and pseudo-labels
- dataset-wide analysis **iteratively** improves image representation
- first study of local activations in few-shot learning
- training to convergence in few-shot learning
- advances on robustness of convolutional networks

part IV

beyond

# outline – part IV

15 current work

16 outlook

# smooth adversarial examples



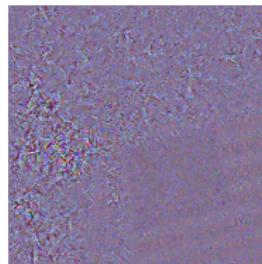
original



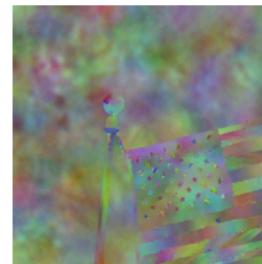
C&W



sC&W



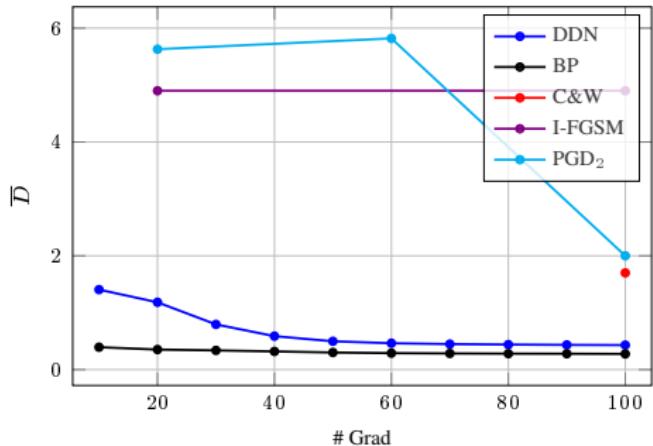
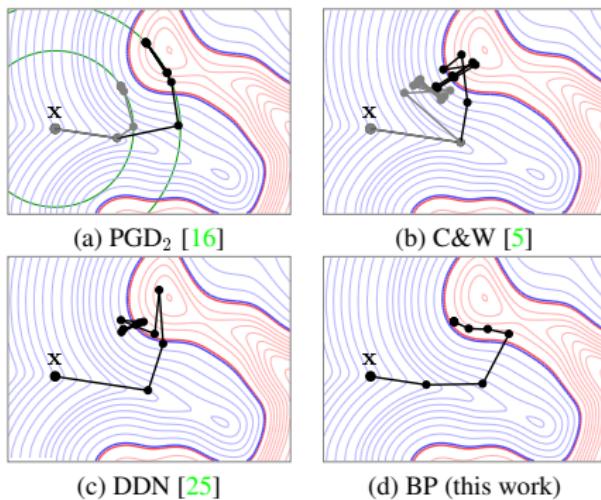
distortion 3.64



distortion 4.59

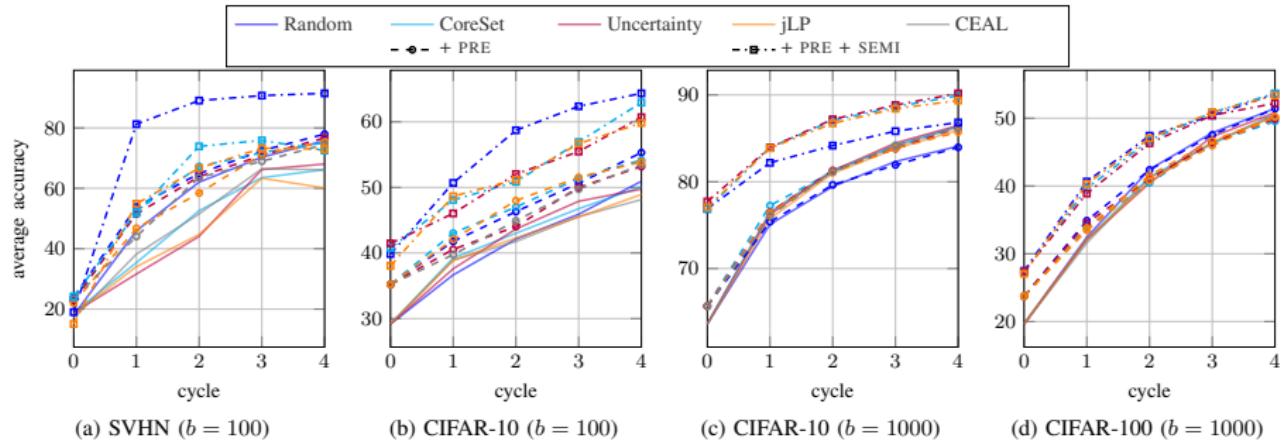
- force perturbation to be 'smooth like' the input image
- despite the extra constraint, the smooth attack performs better

## boundary projection (BP) attack



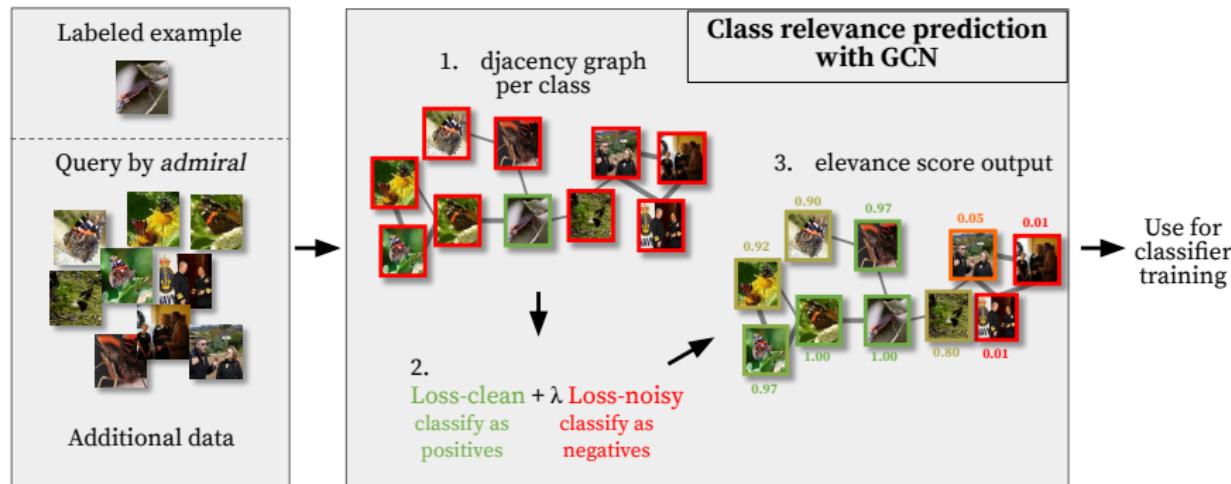
- optimize distortion on class boundary, avoiding oscillations
  - low-distortion adversarial examples at unprecedented speed

# deep active learning



- use **unlabeled data** at model training, not just acquisition
  - surprising improvement, compared to acquisition strategies
  - **random baseline** beats other strategies in low-label regime

## learning from few clean and many noisy labels



- large-scale unlabeled data: YFCC100M
  - graph convolutional network discriminates clean from noisy data

## **few-shot few-shot learning**



- few-shot version of few-shot learning: base class examples are few
  - representation learning on large-scale data of **different domain**
  - **spatial attention** by off-the-shelf ResNet-18 (pre-tained on Places)

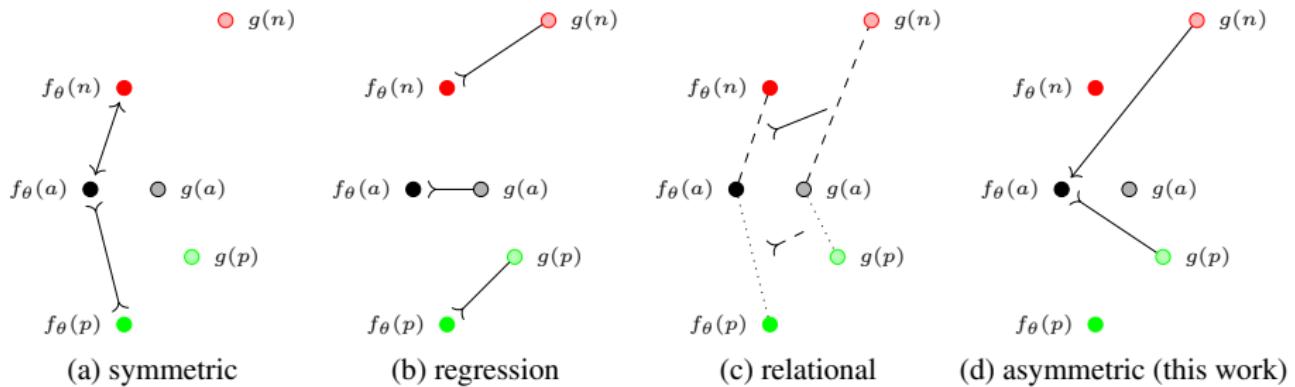
# nano-supervised object detection (NSOD)



- few weakly-labeled and many unlabeled images
- trade off less supervision with more data
- work with unknown classes in the wild

Z. Yang, M. Shi, Y. Avrithis, C. Xu, V. Ferrari. arXiv 2019. Training Object Detectors from Few Weakly-Labeled and Many Unlabeled Images.

## asymmetric metric learning (AML)



- combine supervised metric learning and knowledge transfer
  - compatible with any pair-based loss function
  - EfficientNet-B3 student outperforms ResNet-101 teacher on RevOP

## take home message

**exploring data and learning the representation  
are mutually beneficial**

# outline – part IV

15 current work

16 outlook

# motivation

- computing power still incomparable to biological visual systems
- amount and quality of data still incomparable to what is seen by humans
- human visual long-term memory has a massive capacity
- current architectures are typically stateless

# motivation



- computing power still incomparable to biological visual systems
- amount and quality of data still incomparable to what is seen by humans
- human visual long-term memory has a massive capacity
- current architectures are typically stateless

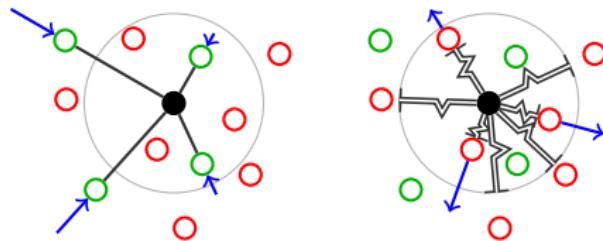
# data as a first-class citizen in visual recognition

- data becomes explicit part of model than just its training process
- translate more **storage capacity** to better performance
- long term goal: **artificial visual long-term memory**

# data as a first-class citizen in visual recognition

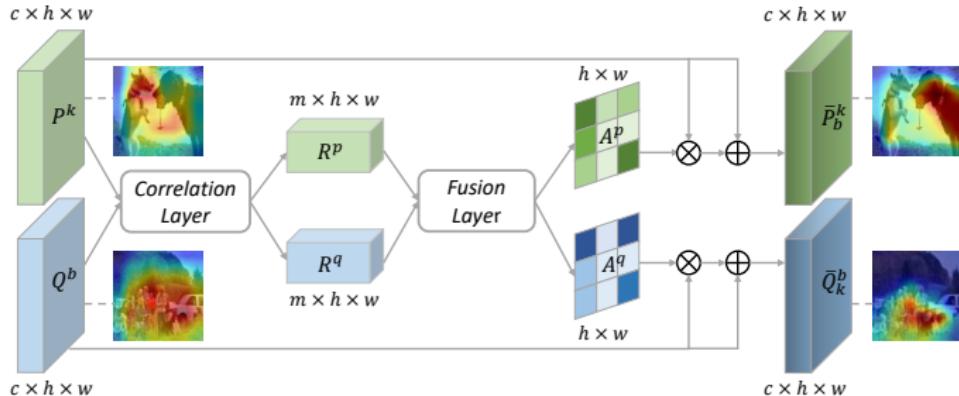
- data becomes explicit part of model than just its training process
- translate more storage capacity to better performance
- long term goal: artificial visual long-term memory

# rethinking metric learning



- unify tasks and loss functions
  - study all **supervision settings** that are common in classification
  - apply loss functions **globally** on the entire dataset
  - extend to detection and instance segmentation

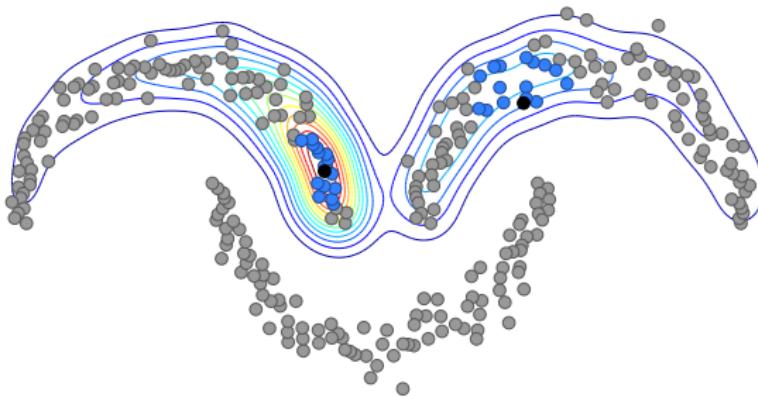
# category-level semantic alignment



- classes represented by tensors
- end-to-end learning using geometric **alignment**
- answer the invariance *vs.* discriminative power dilemma
- encourage **sparse** representations at inference

Hou, Chang, Ma, Shan and Chen. arXiv 2019. Cross Attention Network for Few-shot Classification.  
Siméoni, Avrithis and Chum. CVPR 2019. Local Features and Visual Words Emerge in Activations.

# manifolds, indexing, and geometry

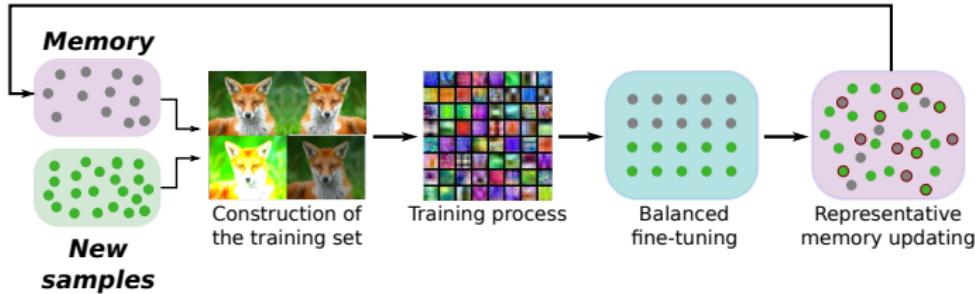


- scale up manifold search to billions
- use **geometry**: extend pairwise affinity from vectors to tensors
- extend to graph convolutional networks

Iscen, Tolias, Avrithis, Furon and Chum. CVPR 2017. Efficient Diffusion on Region Manifolds- Recovering Small Objects with Compact CNN Representations.

Iscen, Tolias, Avrithis, Furon and Chum. CVPR 2018. Fast Spectral Ranking for Similarity Search.

# learning while memorizing



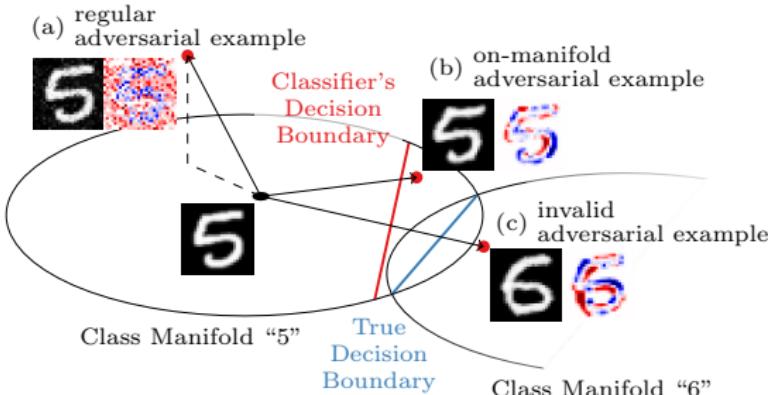
- category-level tasks: a “summary” of training set explicitly memorized
- instance-level tasks: training and test sets become part of a continuously growing knowledge
- memory-based few-shot learning

Lifchitz, Avrithis, Picard and Bursuc. CVPR 2019. Dense Classification and Implanting for Few-Shot Learning.

Iscen, Tolias, Avrithis, Chum, and Schmid. arXiv 2019. Graph convolutional networks for learning with few clean and many noisy labels.

Castro, Marin-Jimenez, Guil, Schmid and Alahari. ECCV 2018. End-to-End Incremental Learning.

# on-manifold adversarial robustness



- adversarial defenses: “ultimate form” of regularization
- hurt on clean data, unless constrained on the manifold (?)
- generalize beyond smoothness and beyond classification
- model the manifold using true data

Stutz, Hein and Schiele. CVPR 2018. Disentangling Adversarial Robustness and Generalization.

Zhang, Avrithis, Furun and Amsaleg. JIS, in press. Smooth Adversarial Examples.

Zhang, Avrithis, Furun, Amsaleg. arXiv 2019. Walking on the Edge: Fast, Low-Distortion Adversarial Examples.



# thank you!

<https://avrithis.net>