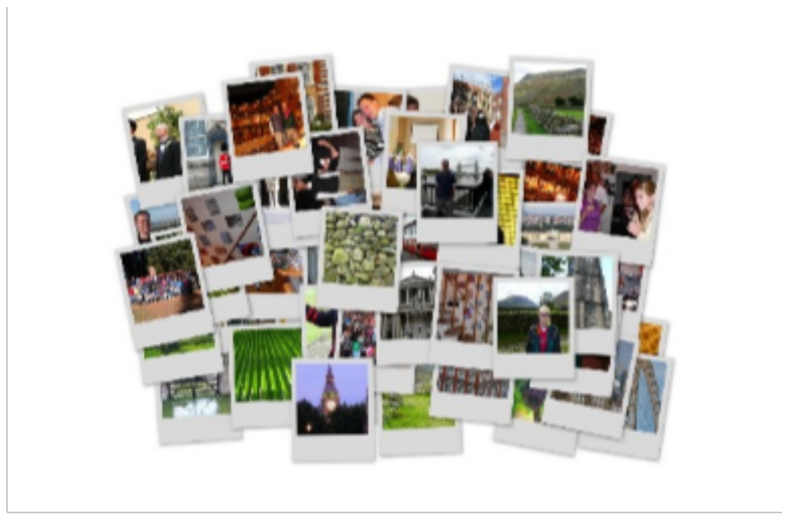
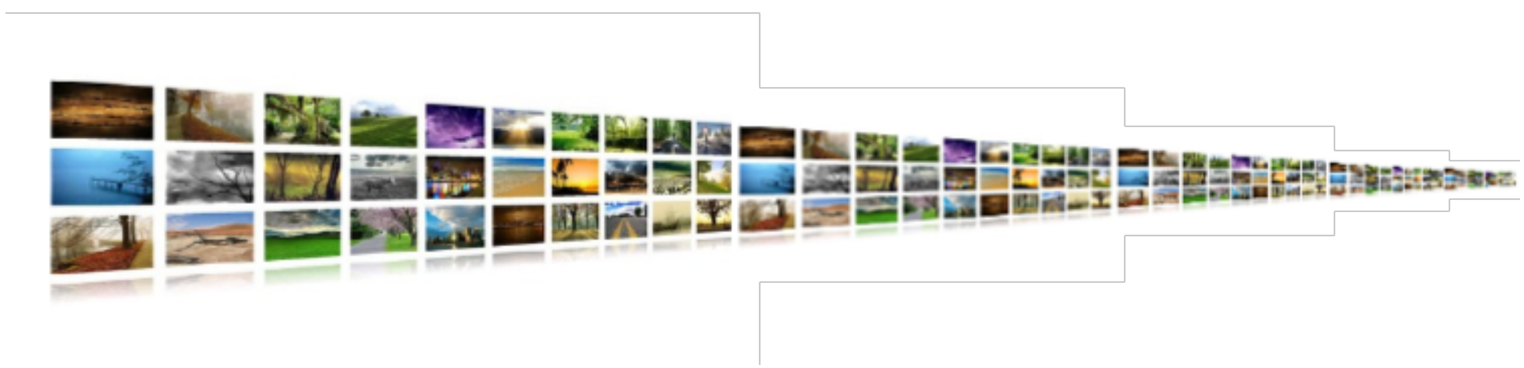
Image Webs

Discovering and using object-manifold structure in large-scale image collections



Kyle Heath - INRIA Saclay Oct 30, 2013

Problem:
Vast collections of images...



Problem:

Vast collections of images...

but virtually no useful metadata

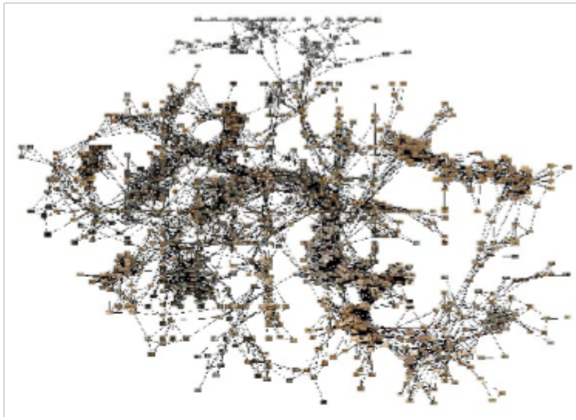
Need automated methods to associate semantic level metadata with images

Approach: Images to object-graphs

Input: millions of images

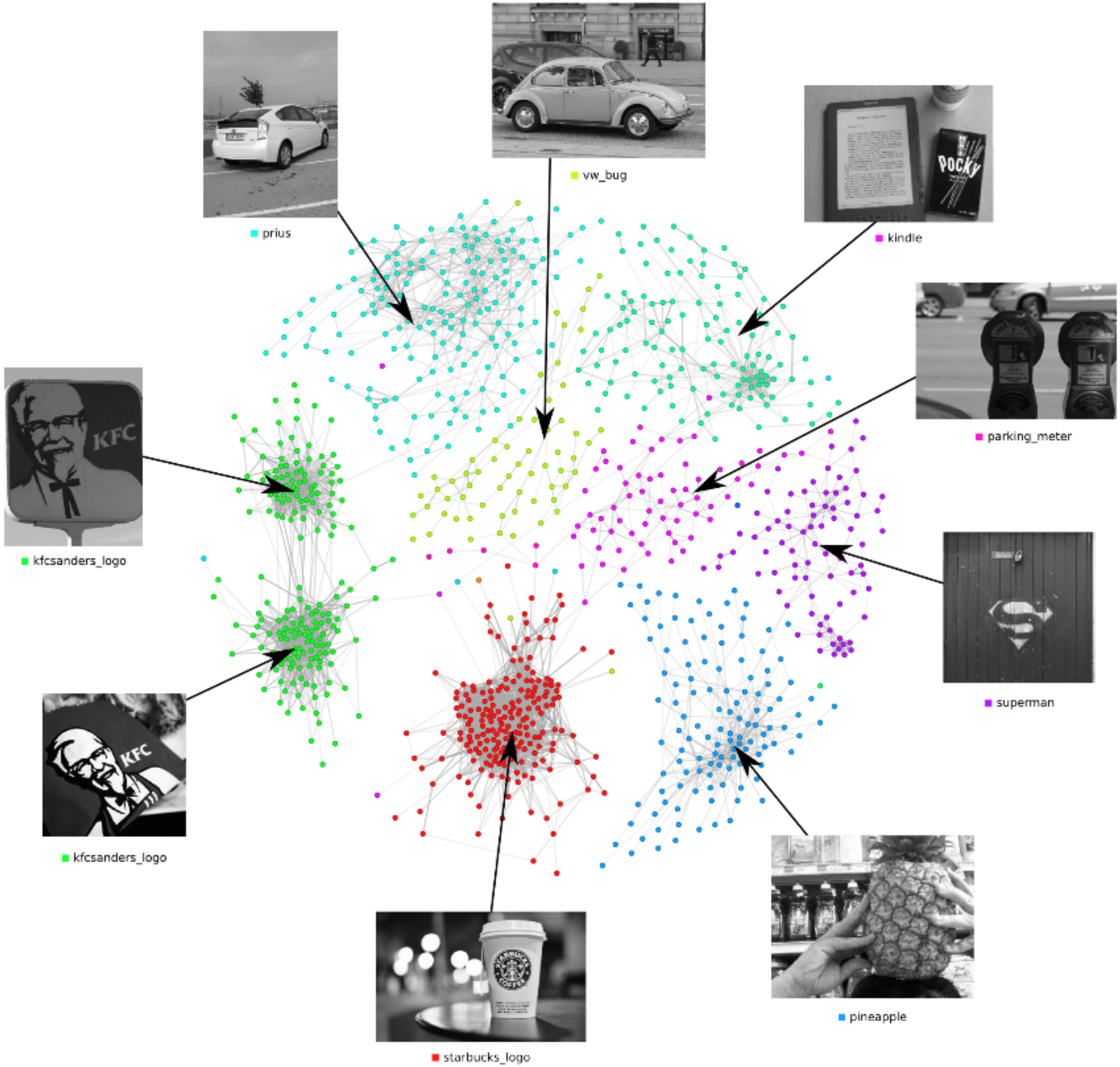
unsupervised

Output: graph encoding "object-instance" appearances



Goal: Construct graphs to approximate the manifold structures induced by objects in images.

An example...


prius
vw_bug
kindle
POCKY
parking_meter
KFC
kfcsanders_logo
superman
KFC
kfcsanders_logo
pineapple
starbucks_logo

Rest of this talk is about...

- building “object” graphs
- using “object” graphs

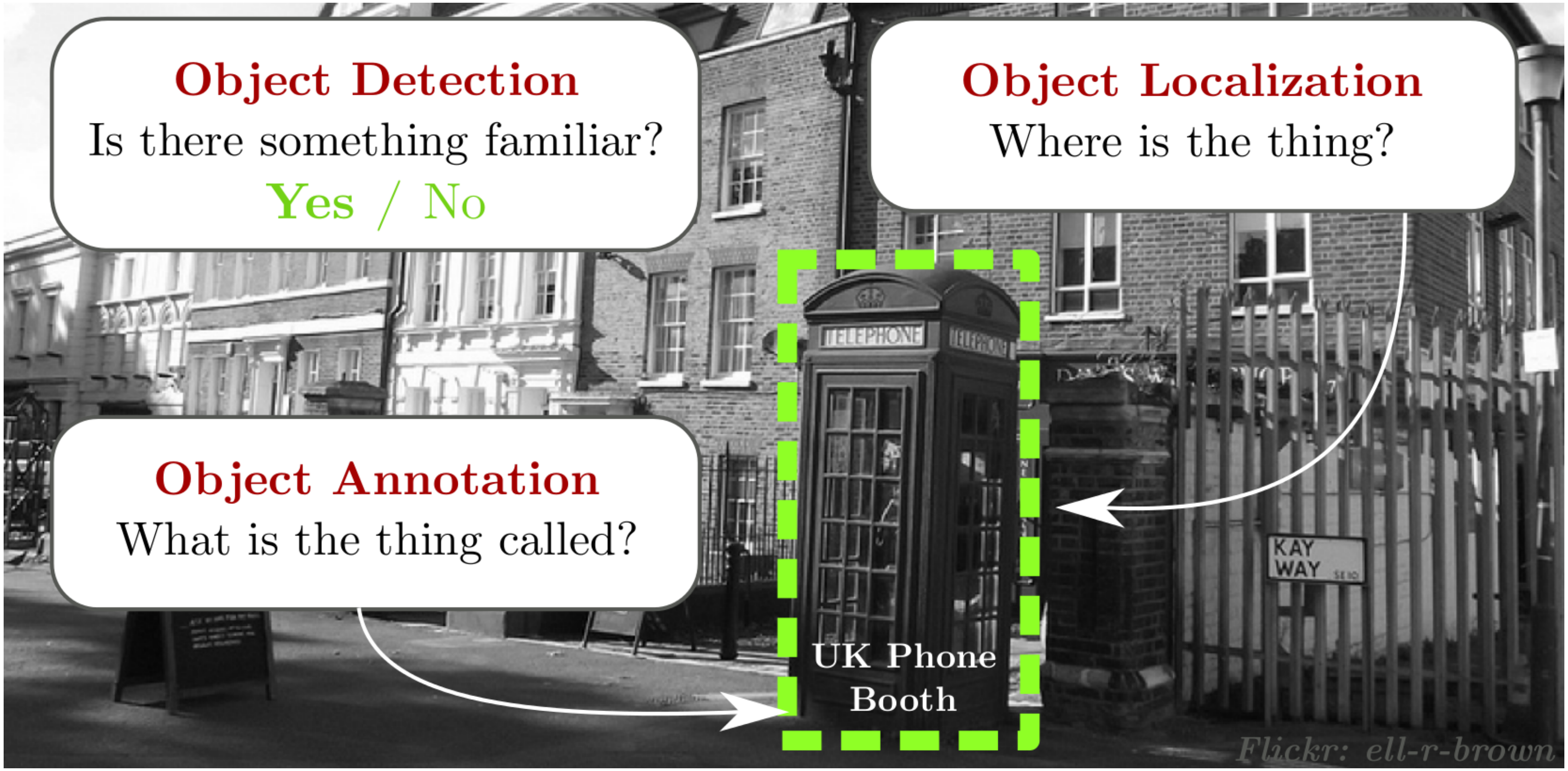
Overview

- Background
- Construction
  - Image-region graph
  - Large-scale image-matching for manifold learning
    Why some local feature matching pipelines are are better for building manifolds
- Applications
  - Fine-grained semi-supervised object recognition
  - Image-collection visualization

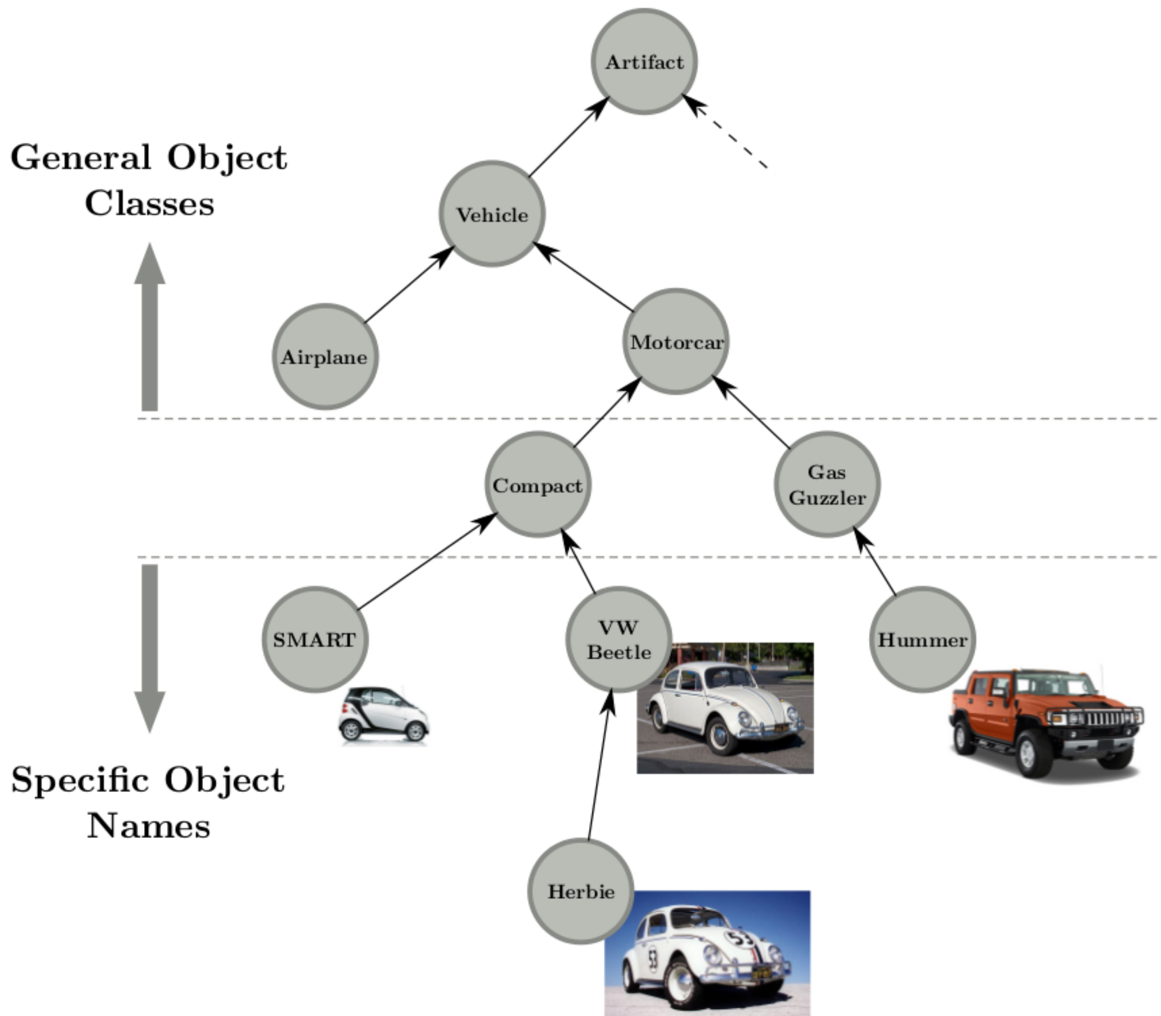
- Bonus Topics
  - Cloud computing for researchers
    When to use the cloud and some tools to make it easier...

Object-recognition has many sub-problems...

here we focus on these three:



Fine-grained object recognition is an open problem



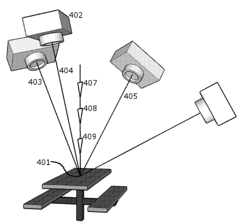
Vision and manifolds

Low-dimensional transformations induce a high-dimensional space of images

Low-Dimension

Camera Motion

6D pose



Object Motion

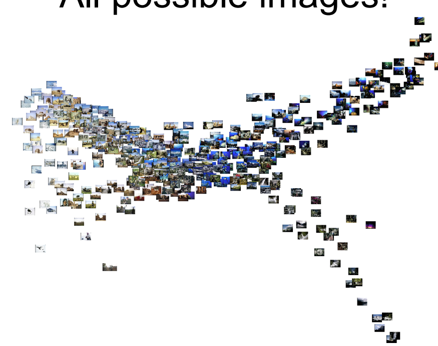
6D pose (+ deformation)

Object Appearance

Color, texture, lighting, variations...

High-Dimension

All possible images!



Vision and manifolds...

Low-dimensional transformations induce a high-dimensional space of images...

Low-Dimension

Camera Motion

6D pose

Object Motion

6D pose (+ deformation)

Object Appearance

Color, texture, lighting, variations...

High-Dimension

Recovering the manifold structure could be useful for many vision problems!

Questions

Given a sample of images, how to discover...

- What objects exist?
- Which objects are related?

Questions

Given a sample of images, how to discover...

- What objects exist?
- Which objects are related?

Q: What is an object?
A: Entities with consistent **geometric structures** that **appear repeatedly** in different scenes but **under variations** in texture, color, lighting, scale / viewpoint.

+ Toothbrush, iPhone, Pineapple

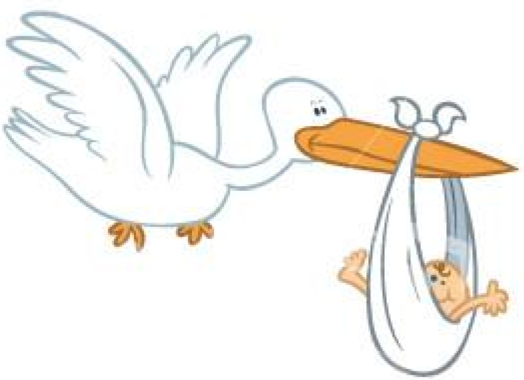
− Sky, Concrete, Water, Grass

Questions

Given a sample of images, how to discover...

- What objects exist?
- Which objects are related?

Q: Where do images come from?
A:

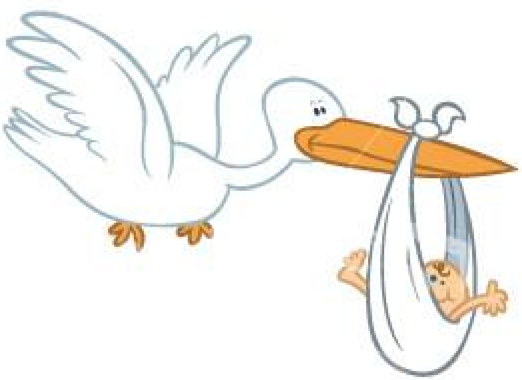


Questions

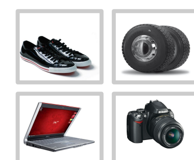
Given a sample of images, how to discover...

- What objects exist?
- Which objects are related?

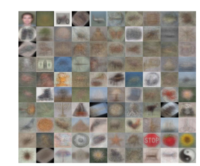
Q: Where do images come from?
A:



Marketing Photos



Web Search



Vacation Photos

Where do images come from?

Proposal: An object-oriented model of image generation... let's call it "P3".

Pick	Place	Perturb
Pick objects from an ontology	Place the objects and the camera in some new scene	Perturb the geometry, color, texture, lighting conditions

Where do images come from?

Pick

Pick objects from an ontology

Telephone

Public Telephone

US

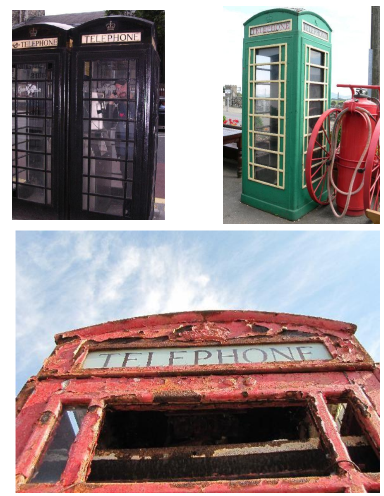
British

Place

Place the objects and the camera in some new scene

Perturb

Perturb the geometry, color, texture, lighting conditions



What can we hope to recover from images?

Pick	Place	Perturb
Pick objects from an ontology	Place the objects and the camera in some new scene	Perturb the geometry, color, texture, lighting conditions

Object Discovery:
What set of objects exist to pick from?

Ontology Learning:
How might objects be related by the "is-a" relation?

Object Context:
Which objects tend to appear together?

Class Variation:
What properties are constant and which vary across instances of the class.

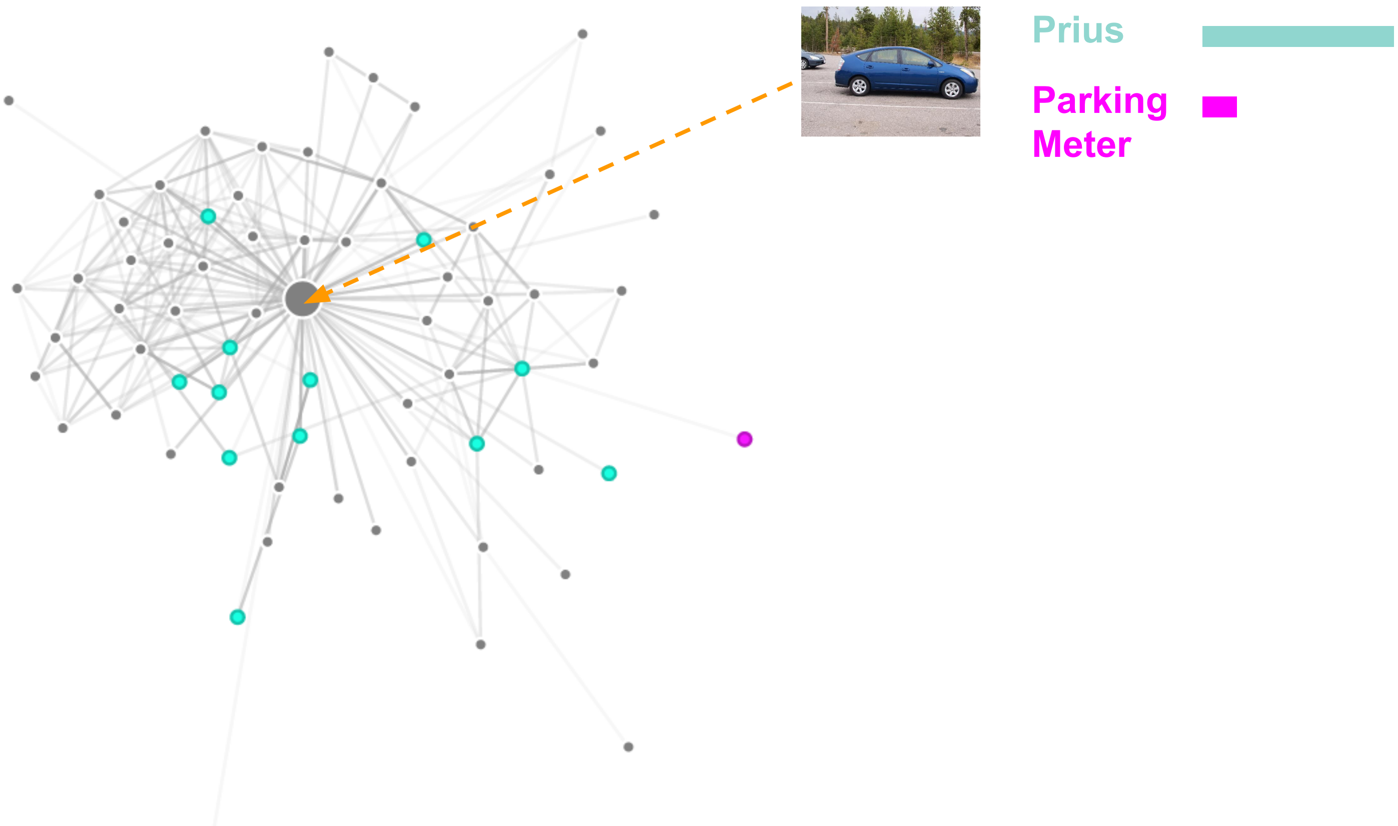
What makes a good graph?

What makes a good graph?

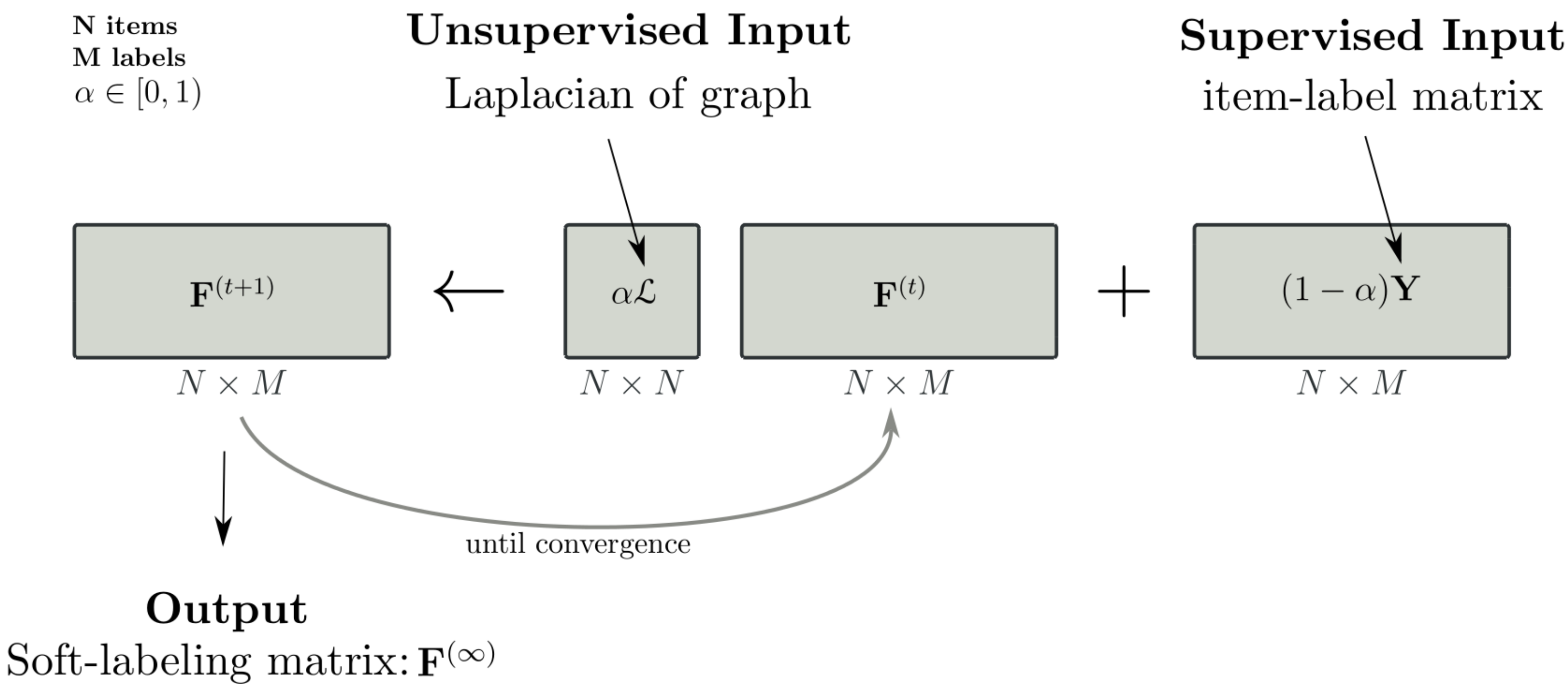
- Clustering = Object Detection

What makes a good graph?

- Label Diffusion = Object Annotation



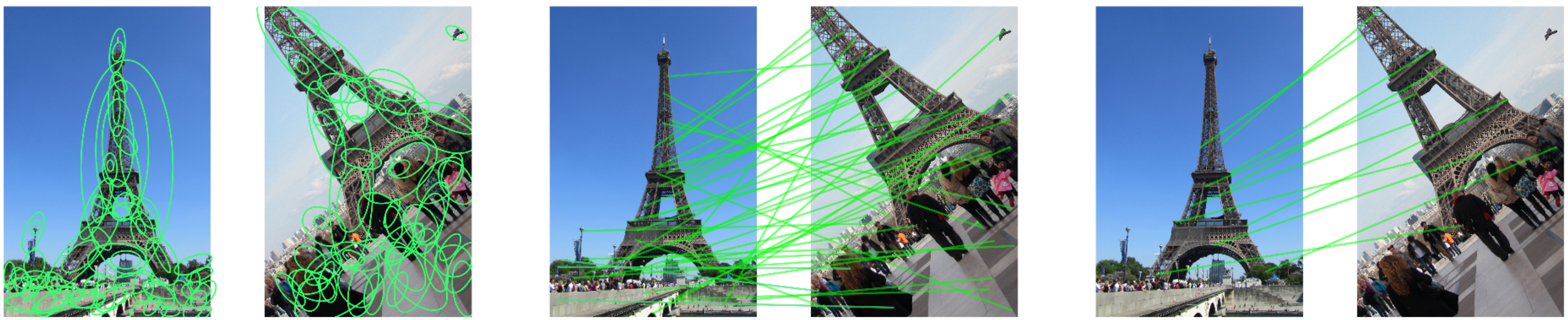
Label propagation



How to build such a graph?

- Local Image Feature Matching

  Detects shared structure under many transformations



(a) Feature extraction (b) Feature matching (c) Geometric verification

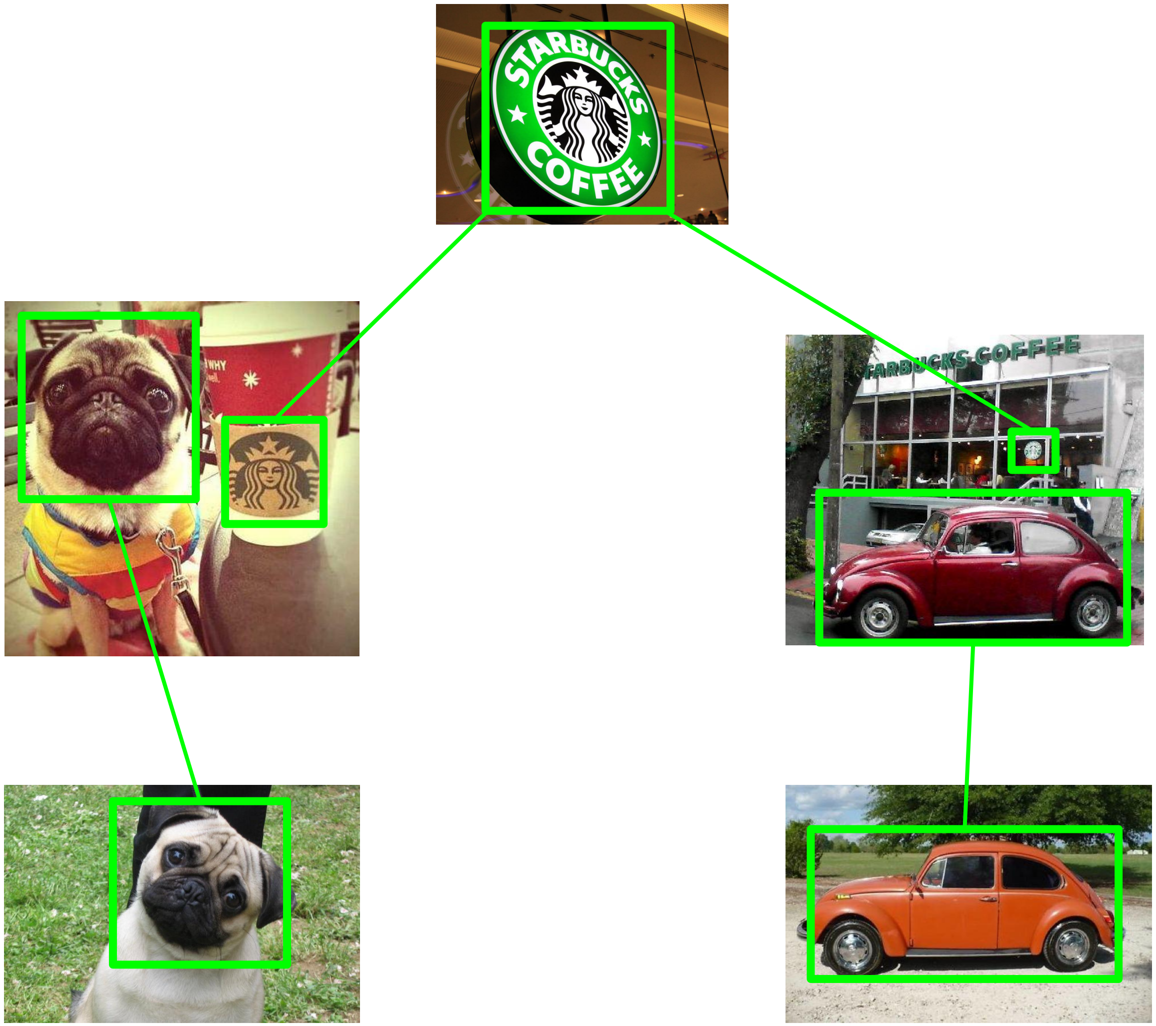
Direct use of Image-Graph can cause label mixing...

Label:
Starbucks Logo

Predicted Label:
Starbucks Logo

Predicted Label:
Starbucks Logo

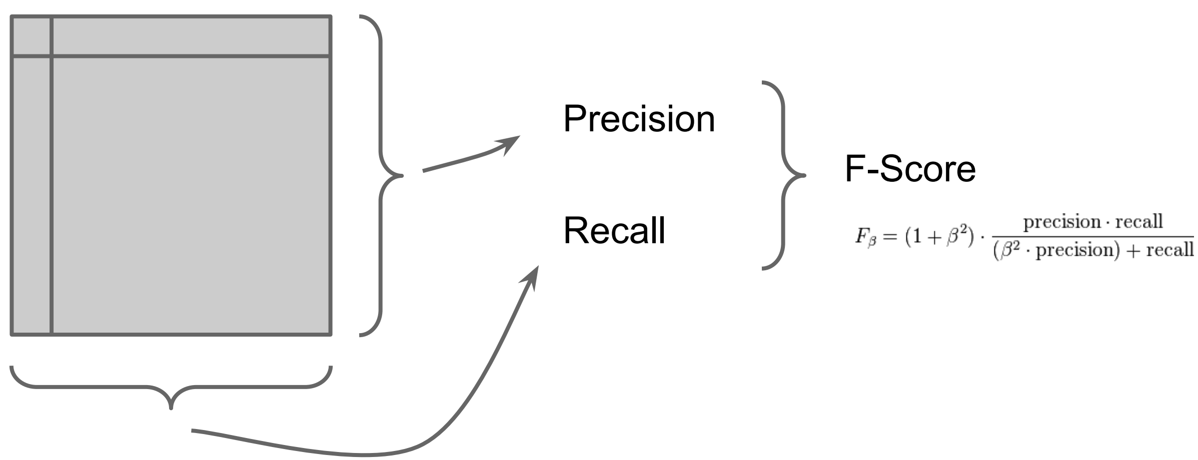
An Image-Region-Graph reduces label mixing...



Evaluation metric

How well does a given graph approximate the true object manifold?

1. Propagate labels on the graph
2. Predict labels for instances with known-labels
   a multi-class classification task
3. Compute confusion-matrix
   and related metrics precision, recall, f-score

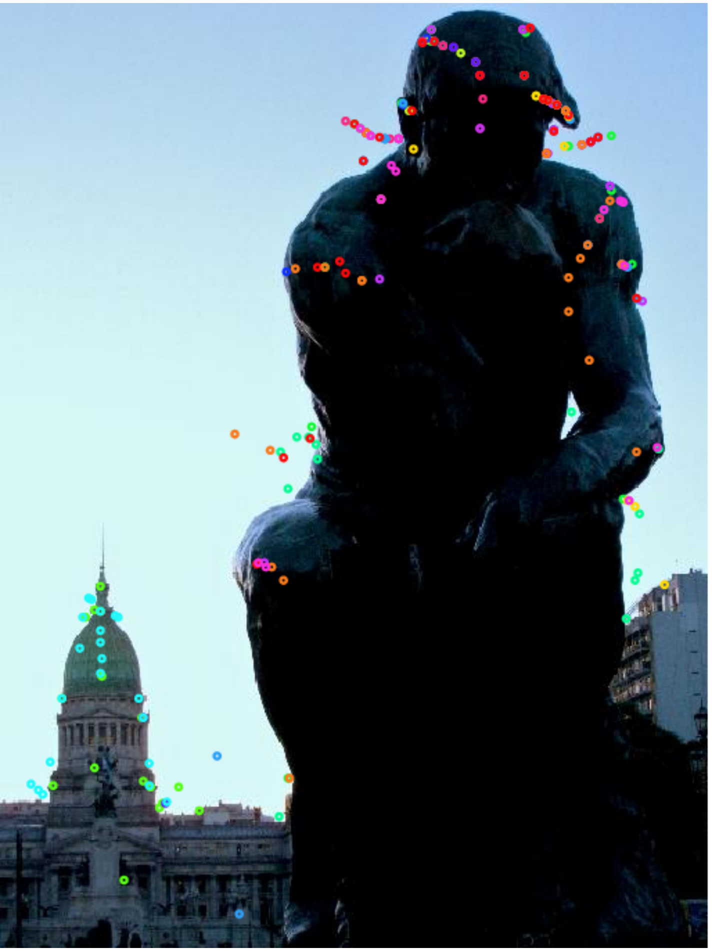


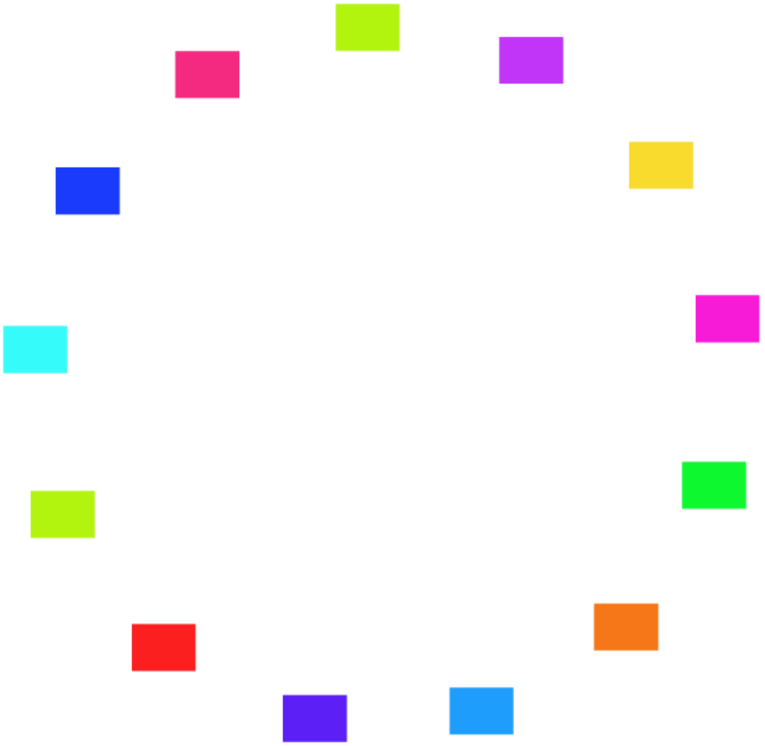
Review

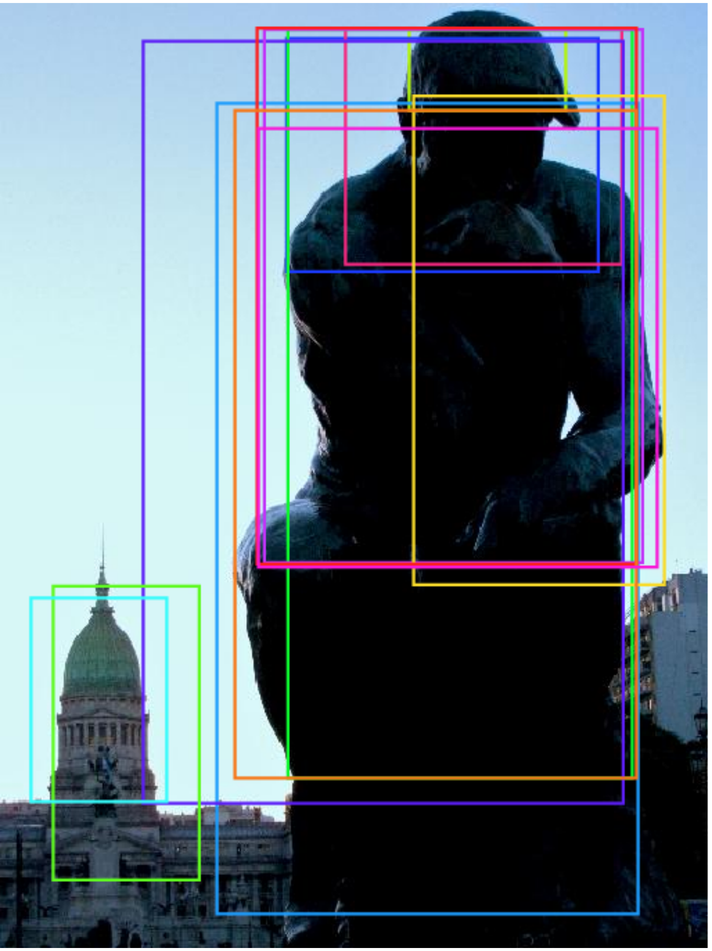
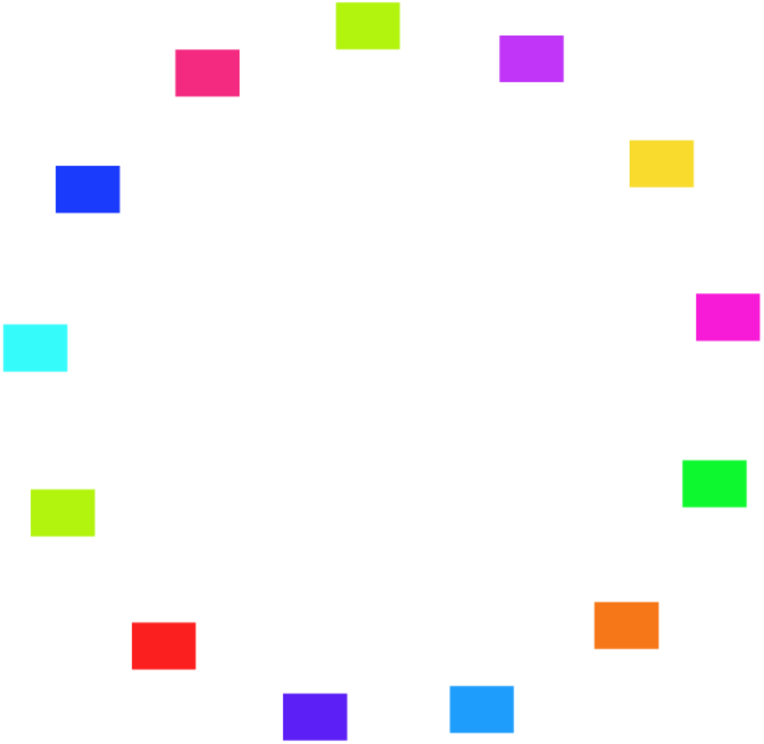
- Image Region Graph
  - Approximates object manifolds
- Applications
  - Object detection, localization, segmentation
    Given just image pixels
  - Soft-classification, image annotation
    Given many image pixels and very little metadata
- Evaluation metrics
  - Label propagation
    Precision, Recall, F-Score

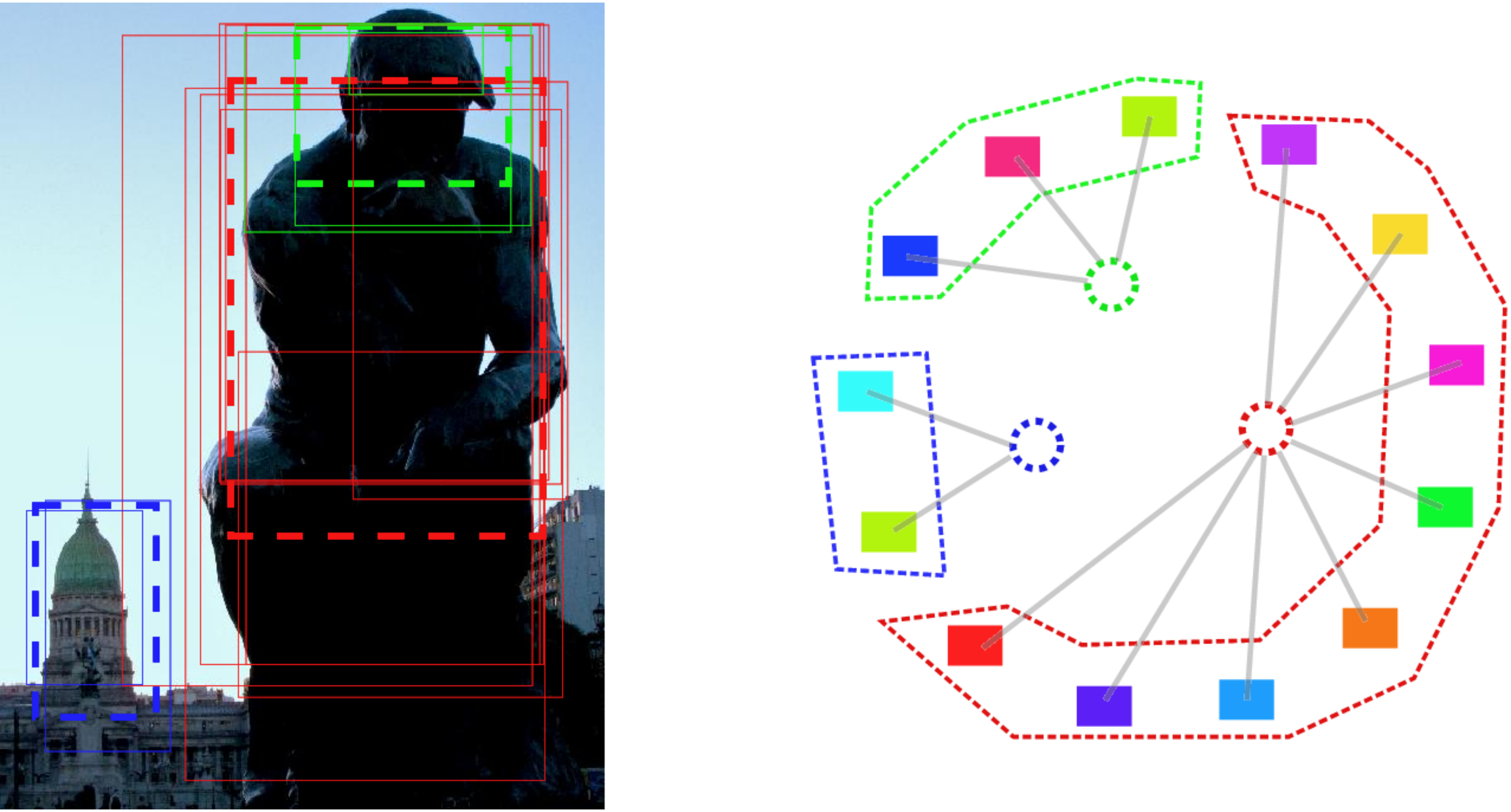
Clustered Image Region Graph Construction

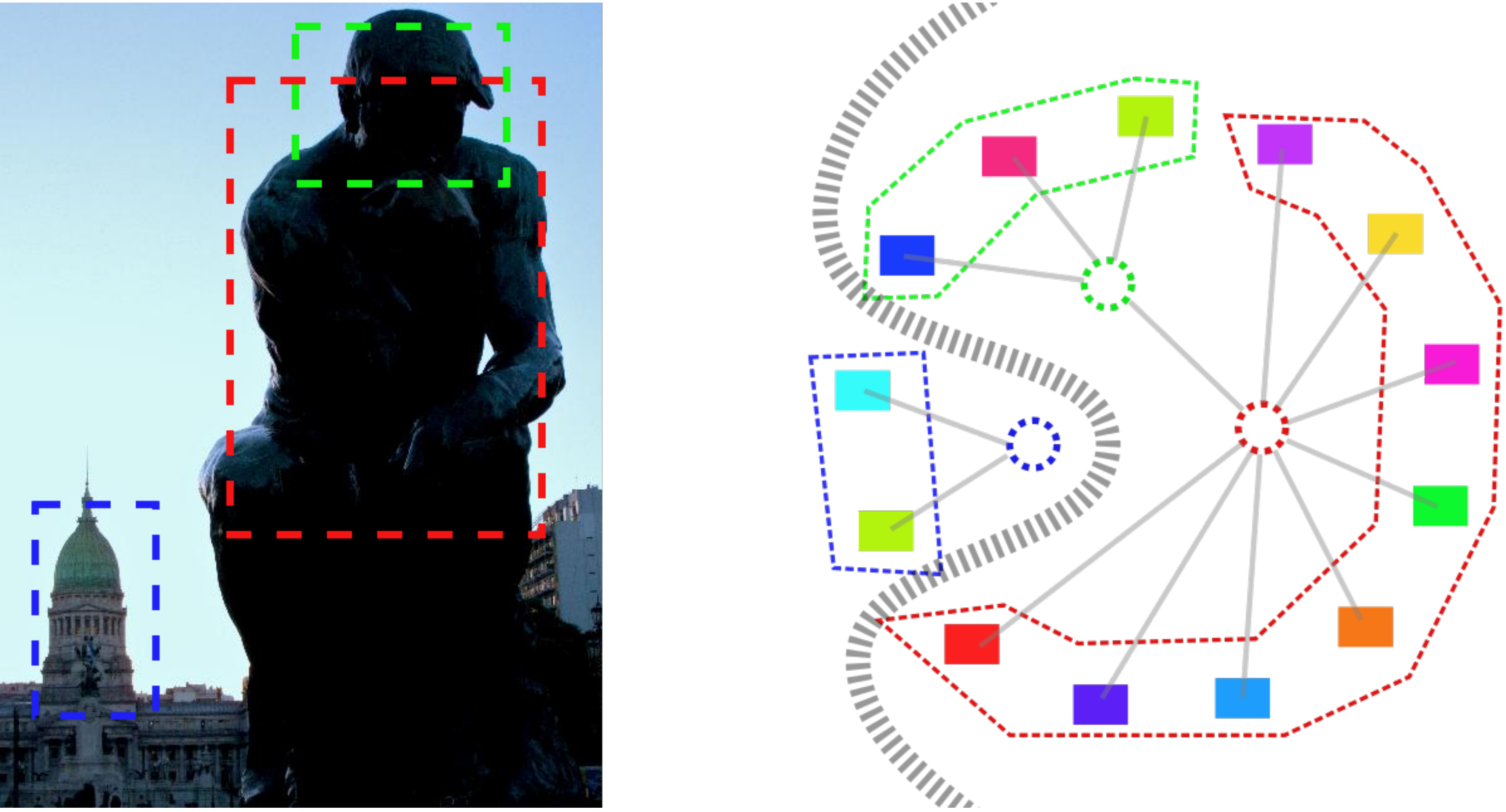
Example follows...







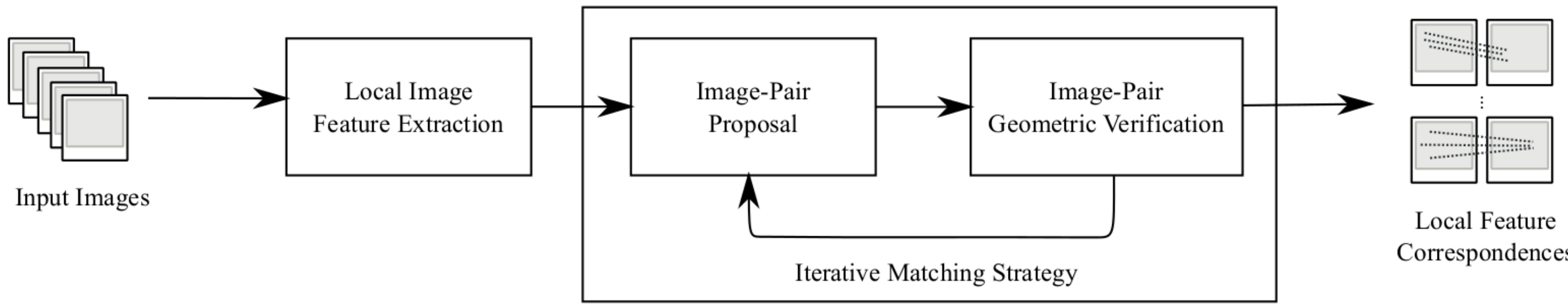


Clustered Image Region Graph

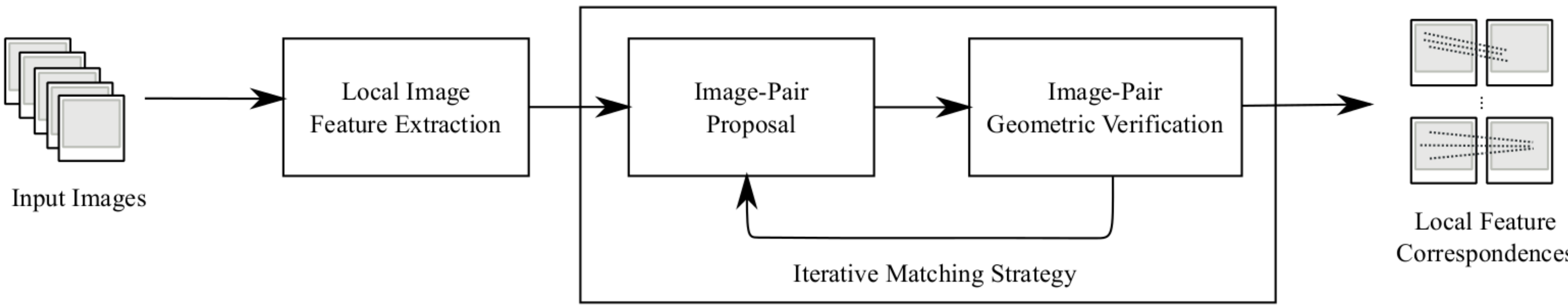
For algorithm, see [thesis]

For code, see [github]

Design space: Local image-feature matching pipelines

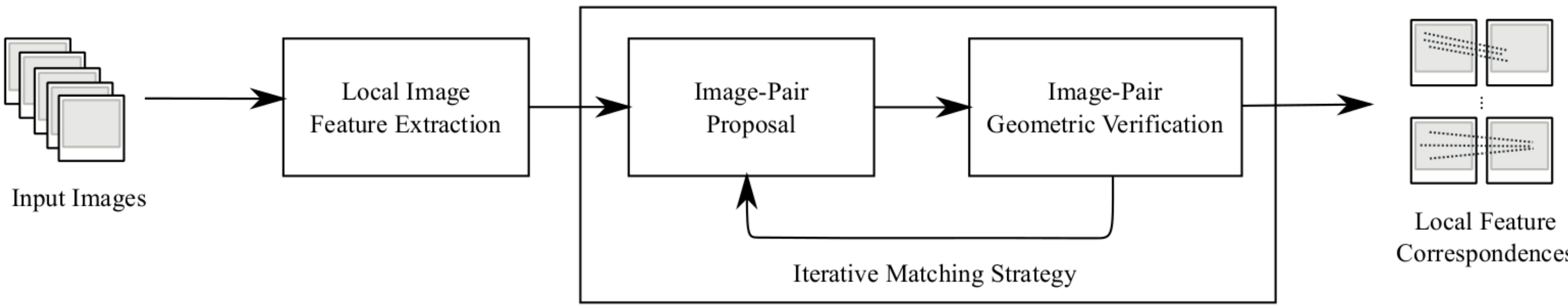


Design space: Local image-feature matching pipelines



Which feature-matching techniques best capture the local metric structure of the object-manifold?

Design space: Local Image-feature matching pipelines



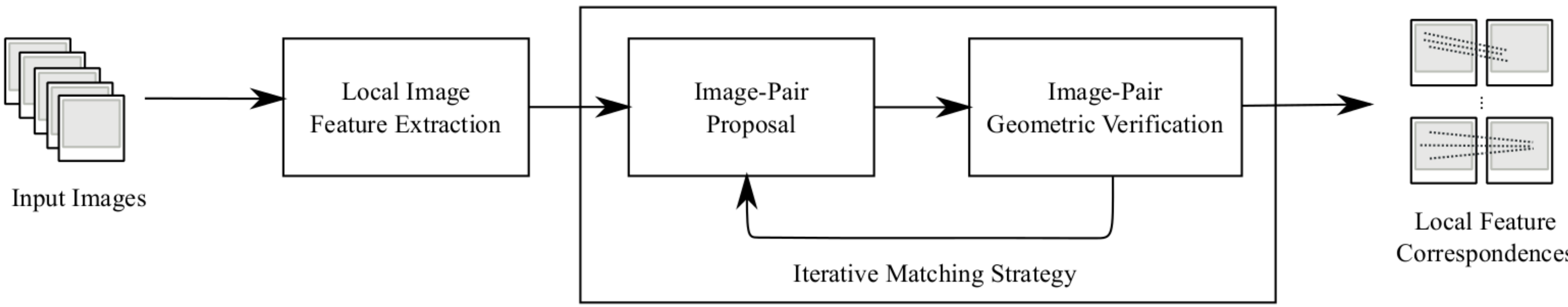
Keypoint Invariance

- Scale + Trans
- Scale + Trans + Rotation
- Affine

Content Based Image Retrieval

- Full Representation (ANN)
- Bag-Of-Words (Quantization)

Design space: Local Image-feature matching pipelines



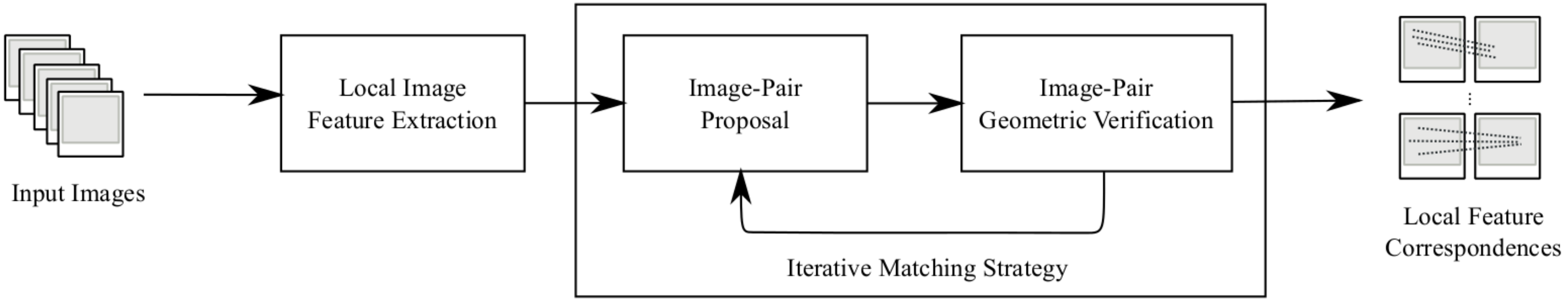
Keypoint Invariance

- Scale + Trans
- Scale + Trans + Rotation
- Affine

Content Based Image Retrieval

- Full Representation (ANN)
- Bag-Of-Words (Quantization)

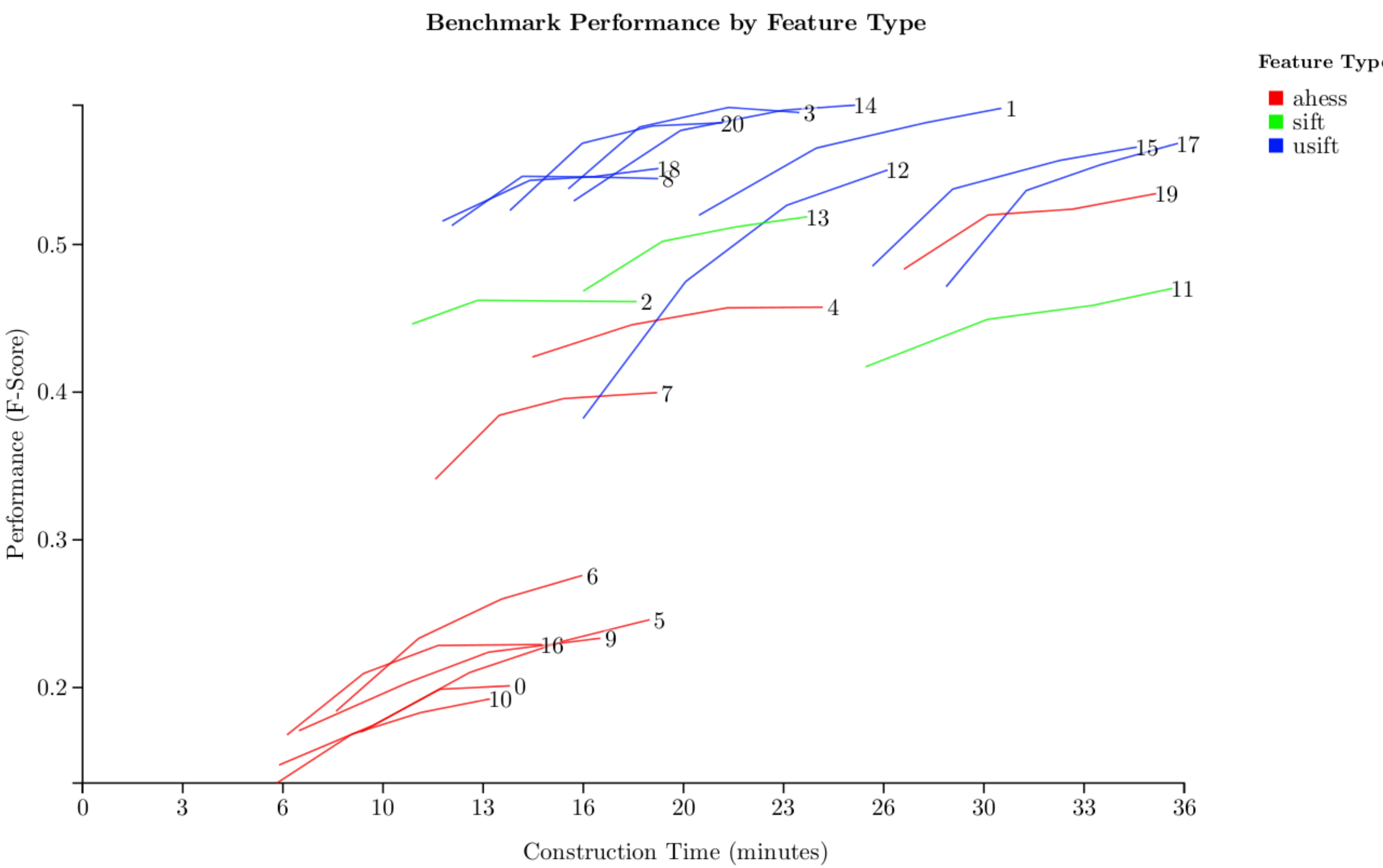
Design space: Local Image-feature matching pipelines



Keypoint Invariance
- Scale + Trans
- Scale + Trans + Rotation
- Affine

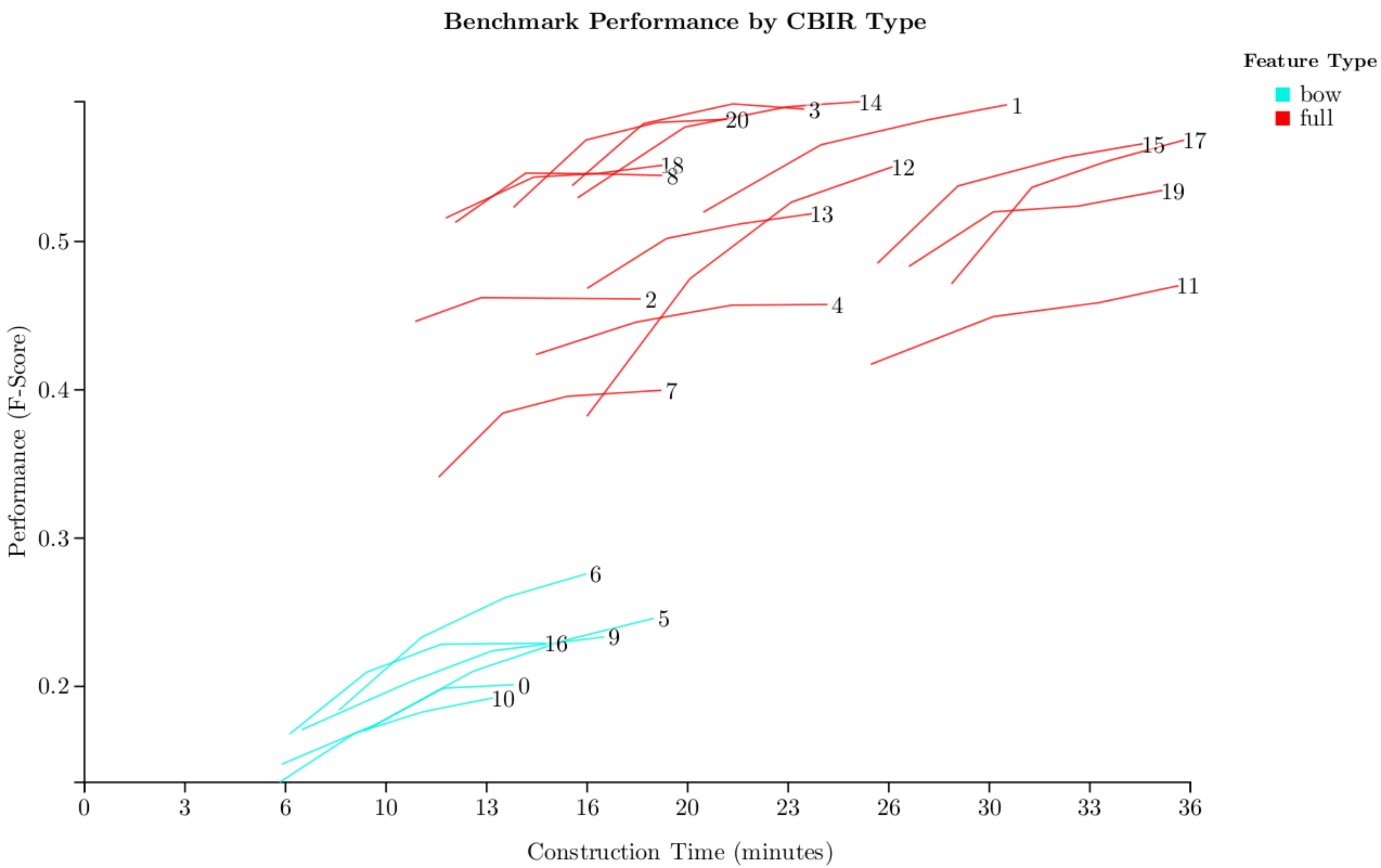
Content Based Image Retrieval
- Full Representation (ANN)
- Bag-Of-Words (Quantization)

Which local keypoint type is best? *



* For the task of approximating the local metric structure of object-manifolds

Which CBIR method is best? *



* For the task of approximating the local metric structure of object-manifolds

Measuring end-to-end performance is important

Observations *(perhaps surprising)*

- Simpler keypoint detection actually better
  3 DOF > 4 DOF > 6 DOF
- Full-Representation CBIR better than Bag-Of-Words CBIR
  Much better results in similar runtime

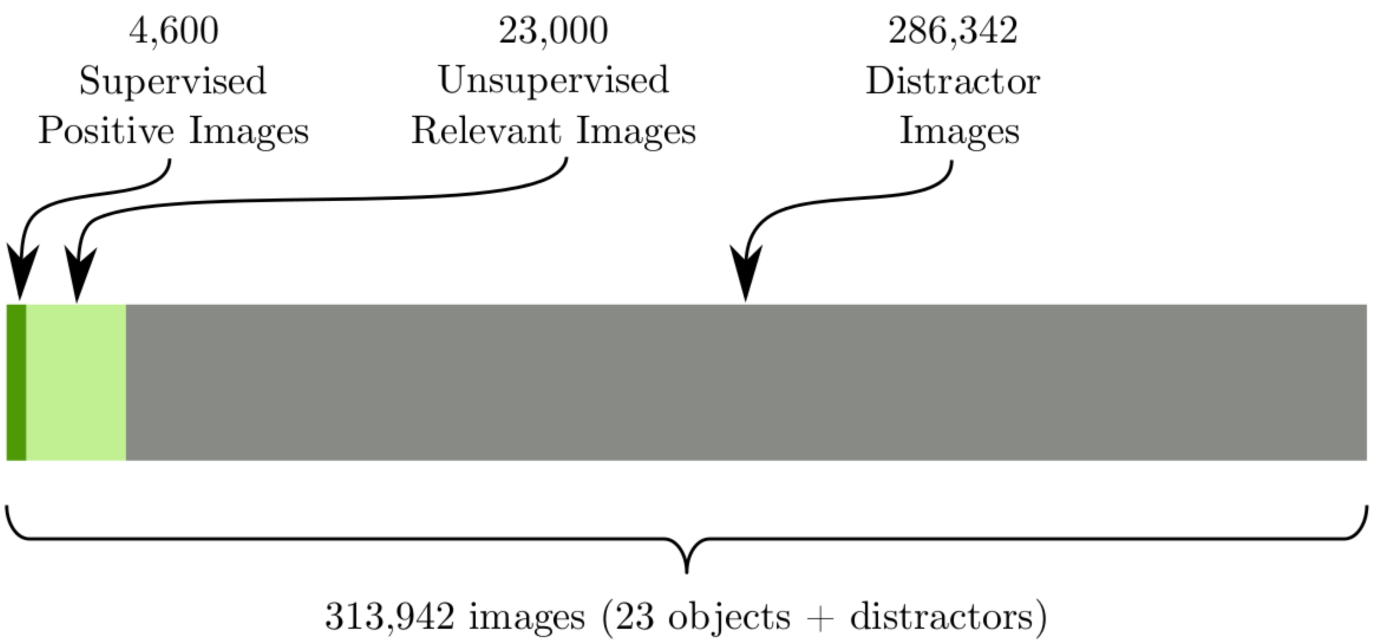
Applications

- Fine-grained semi-supervised object recognition
- Image-collection visualization

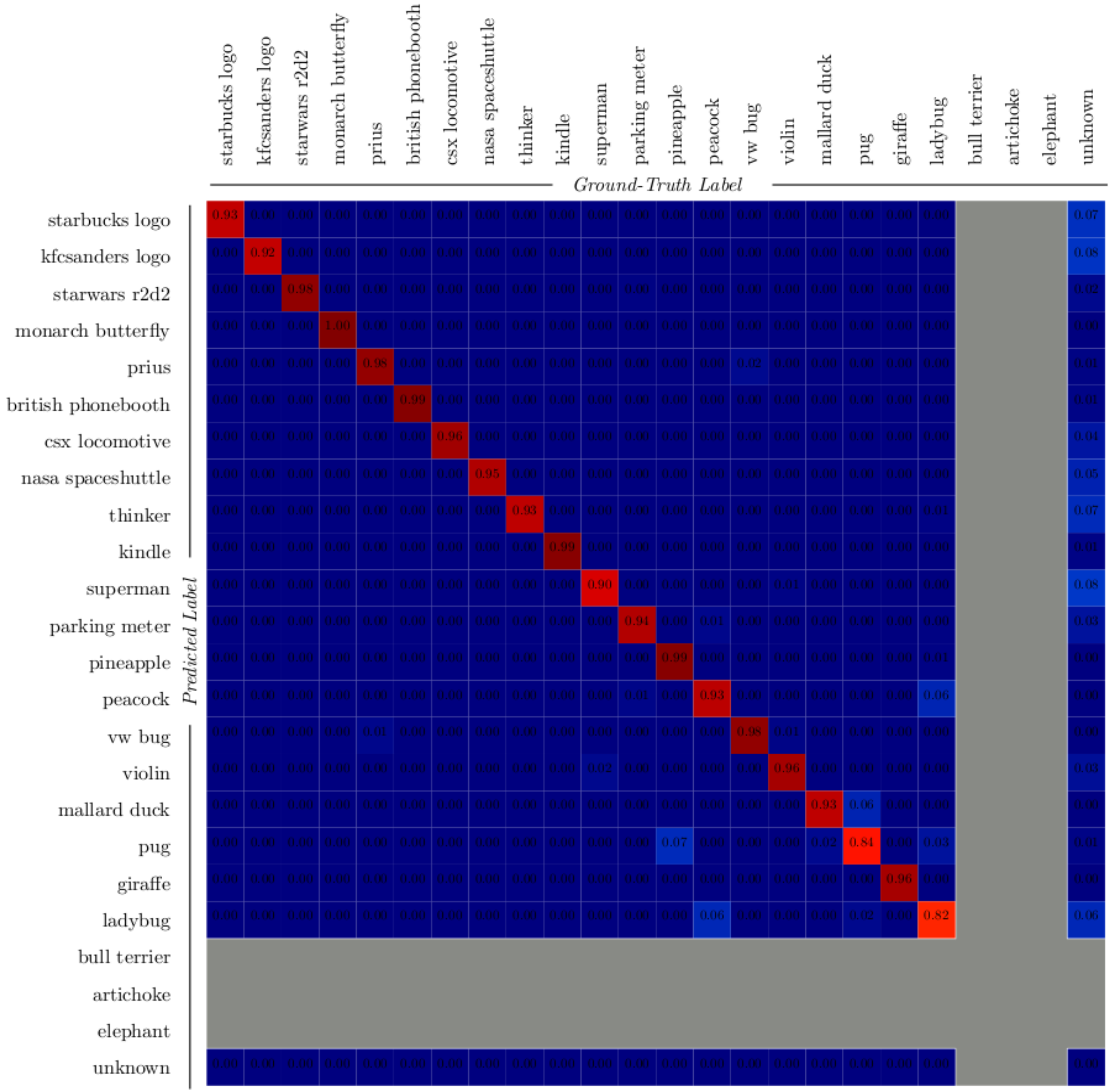
Fine-grained semi-supervised object recognition

Dataset: TIDE

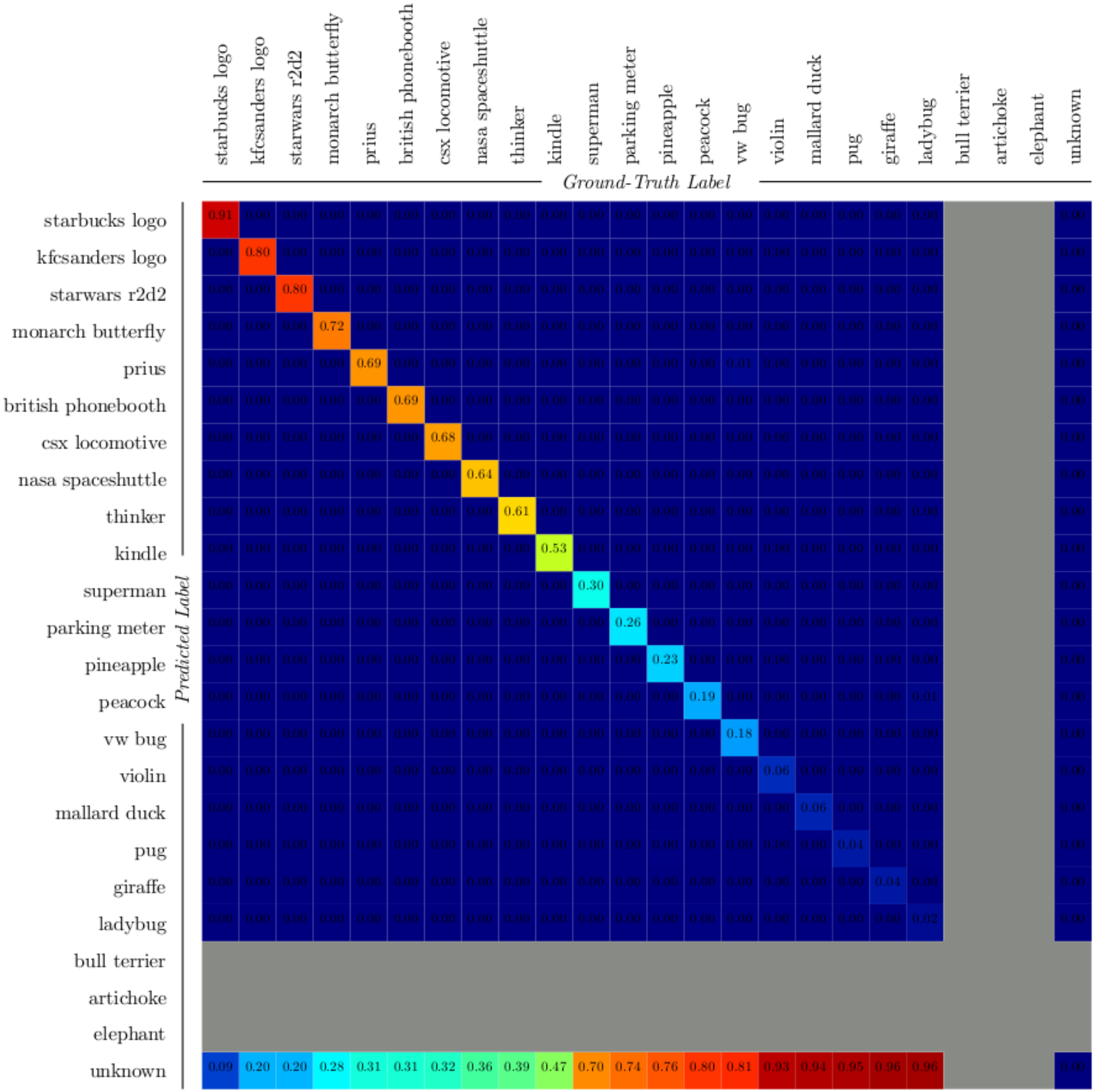
Dataset: TIDE+Holiday



Precision


Ground-Truth Label
Predicted Label
starbucks logo
kfcsanders logo
starwars r2d2
monarch butterfly
prius
british phonebooth
csx locomotive
nasa spaceshuttle
thinker
kindle
superman
parking meter
pineapple
peacock
vw bug
violin
mallard duck
pug
giraffe
ladybug
bull terrier
artichoke
elephant
unknown

Recall


Ground-Truth Label
Predicted Label
starbucks logo
kfcsanders logo
starwars r2d2
monarch butterfly
prius
british phonebooth
csx locomotive
nasa spaceshuttle
thinker
kindle
superman
parking meter
pineapple
peacock
vw bug
violin
mallard duck
pug
giraffe
ladybug
bull terrier
artichoke
elephant
unknown
0.91
0.80
0.80
0.72
0.69
0.69
0.68
0.64
0.61
0.53
0.30
0.26
0.23
0.19
0.18
0.06
0.06
0.04
0.04
0.02
0.09 0.20 0.20 0.28 0.31 0.31 0.32 0.36 0.39 0.47 0.70 0.74 0.76 0.80 0.81 0.93 0.94 0.95 0.96 0.96
0.00

Rigid objects are easier...

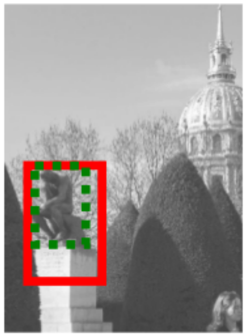
Object	Precision	Recall	F-Score
starbucks logo	0.914	0.892	0.903
starwars r2d2	0.977	0.781	0.868
kfcsanders logo	0.915	0.788	0.847
UK phonebooth	0.996	0.689	0.814
prius	0.969	0.676	0.797
csx locomotive	0.952	0.654	0.776
nasa spaceshuttle	0.947	0.629	0.756
thinker	0.937	0.622	0.748
kindle	0.982	0.500	0.663
superman	0.911	0.301	0.453
parking meter	0.953	0.221	0.358
vw bug	0.959	0.178	0.300
violin	0.965	0.068	0.128
mean	0.952	0.538	0.647

(a) Rigid objects

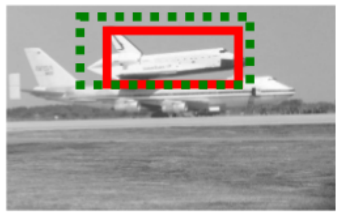
Object	Precision	Recall	F-Score
monarch butterfly	0.997	0.715	0.833
pineapple	1.000	0.227	0.370
peacock	0.961	0.190	0.317
giraffe	0.902	0.044	0.083
pug	0.893	0.042	0.080
mallard duck	0.935	0.039	0.074
ladybug	0.660	0.021	0.040
bull terrier	0.533	0.004	0.007
artichoke	1.000	0.002	0.003
elephant	0.800	0.000	0.000
mean	0.868	0.128	0.181

(b) Non-rigid objects

Success!

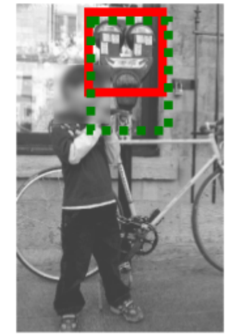


Score	Label
5.4e-04	thinker
1.2e-10	bull_terrier
2.9e-12	violin



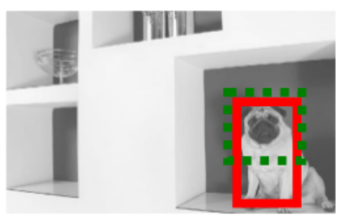
Score	Label
1.7e-03	nasa_spaceshuttle
5.7e-11	pineapple
7.4e-13	ladybug

Score	Label
1.3e-03	monarch_butterfly
9.4e-12	pineapple
3.7e-13	thinker

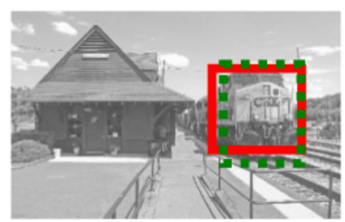


Score	Label
1.3e-03	parking_meter

Score	Label
4.6e-03	mallard_duck



Score	Label
7.6e-03	pug



Score	Label
3.3e-04	csx_locomotive
8.6e-13	pug
1.2e-13	giraffe

Score	Label
1.2e-04	pineapple
3.2e-10	vw_bug
1.5e-10	ladybug

Fail:
Confused with similar object

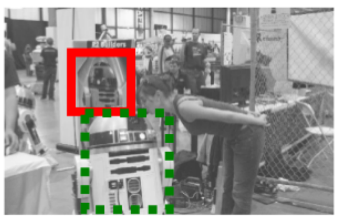
Score	Label
2.0e-04	prius
1.7e-09	vw_bug

Score	Label
2.6e-04	prius
7.8e-06	vw_bug
2.1e-10	parking_meter

Score	Label
1.1e-03	vw_bug
5.8e-06	prius
6.4e-12	elephant

Score	Label
8.4e-04	vw_bug
8.4e-06	prius
1.2e-11	parking_meter

Fail:
Incomplete ground truth



Score	Label
1.2e-05	starwars_r2d2
3.5e-12	british_phonebooth
7.2e-13	parking_meter
5.6e-15	bull_terrier

Score	Label
2.7e-06	starbucks_logo
5.2e-13	parking_meter

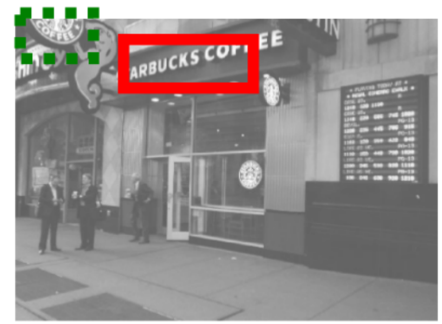
Score	Label
2.3e-04	starbucks_logo
6.3e-10	bull_terrier
2.1e-10	vw_bug
1.8e-11	parking_meter

Score	Label
3.7e-06	kfcsanders_logo

Fail:
Leaking labels

Score	Label
1.0e-04	thinker
1.3e-14	kindle

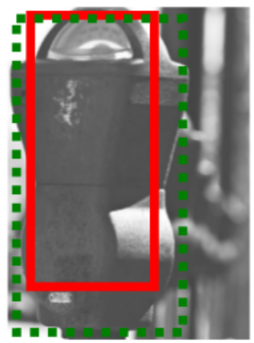
Score	Label
3.4e-04	kfcsanders_logo
3.7e-14	giraffe
1.3e-14	starwars r2d2



Score	Label
1.0e-03	starbucks_logo
2.8e-10	vw_bug
4.9e-12	parking meter

Fail:
Localized... but wrong label

Score	Label
3.0e-06	ladybug
4.4e-13	mallard_duck



Score	Label
1.7e-03	peacock
1.9e-10	violin
9.9e-11	thinker
9.5e-11	starbucks_logo
4.2e-11	parking_meter
2.6e-11	pineapple

Score	Label
5.1e-04	violin
3.6e-06	superman
1.9e-09	pug

Score	Label
2.3e-06	mallard_duck
2.0e-07	pineapple
2.6e-12	violin
1.7e-12	parking_meter
1.4e-12	ladybug
7.8e-13	vw_bug

Is this performance good?

Baseline Comparison: State-of-the-art texture features + SVM by Pintos*

	Object Manifold		SVM	
Type	Precision	Recall	Precision	Recall
rigid	**0.85**	**0.54**	0.52	0.48
non-rigid	**0.87**	0.12	0.42	**0.39**
	Transductive, Localization + Classification		*Discriminative, classification only*	

Yes: much higher precision across all objects (though lower recall on non-rigid objects)

** Nicolas Pinto, David D Cox, and James J DiCarlo. Why is real-world visual object recognition hard? PLoS computational biology, 4(1):e27, 2008.*

Is the region-graph better than just image-graph?

Graph Type	Precision	Recall
Image	0.965	0.643
Image-Region	**1.00**	**0.733**

(a) Cluttered logos dataset

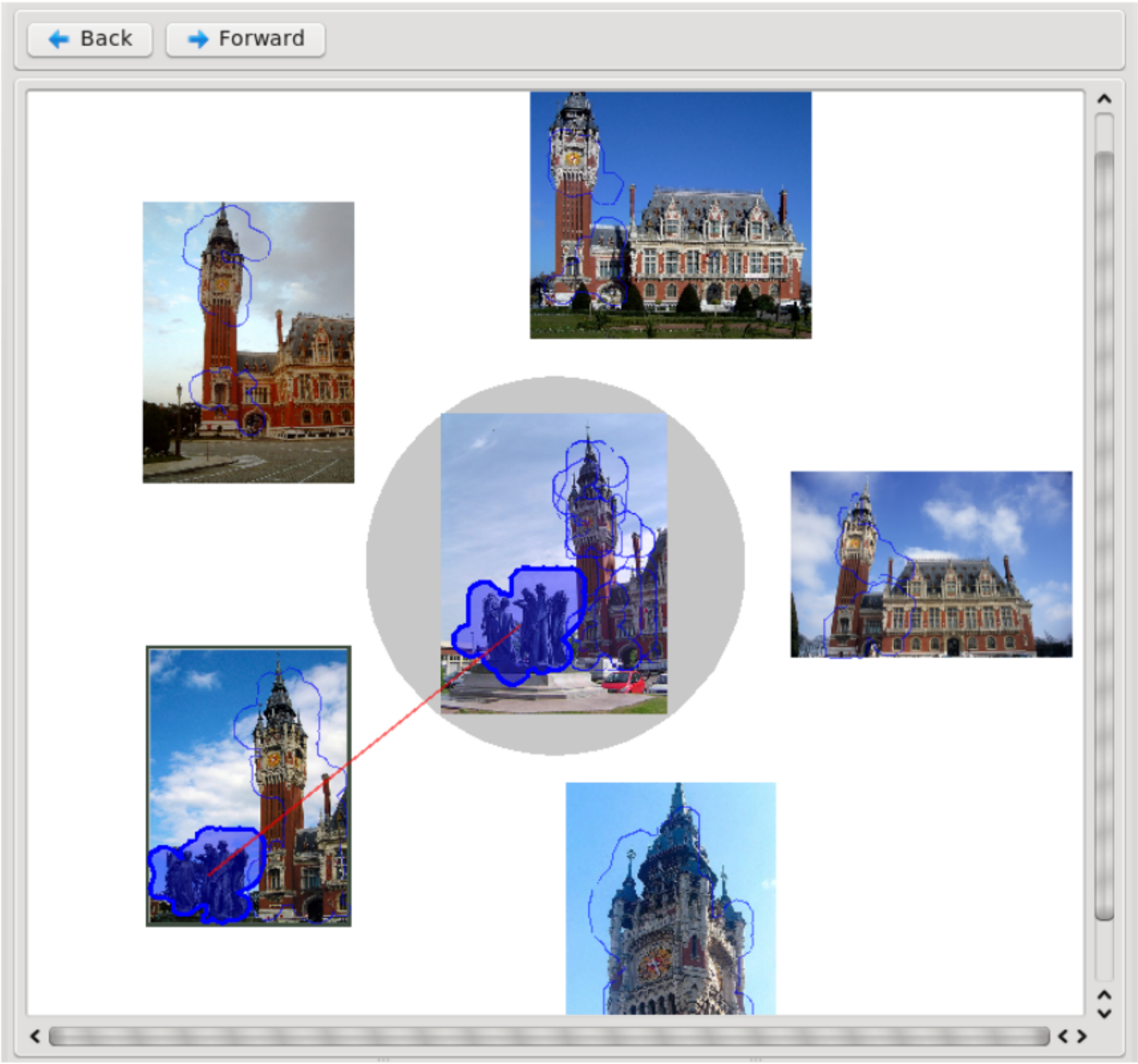
Graph Type	Precision	Recall
Image	0.892	**0.284**
Image-Region	**1.00**	0.271

(b) Cluttered cars dataset

Yes: higher precision, and similar or better recall

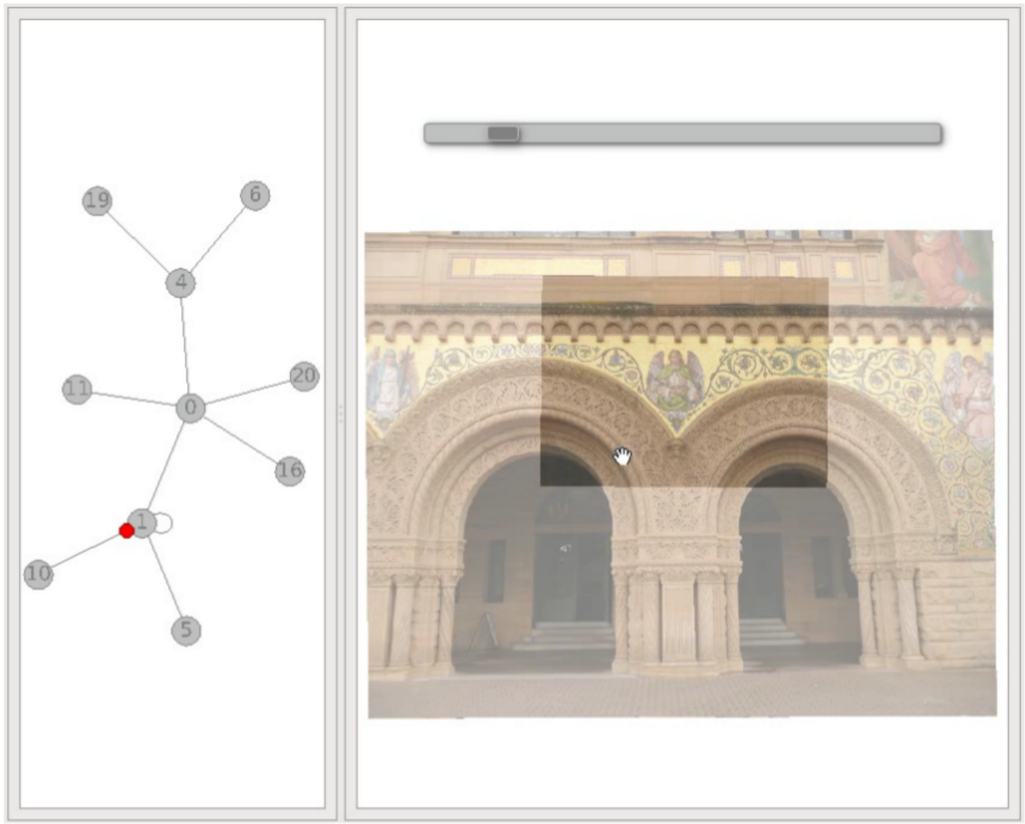
Image-collection visualization

Visual-hyperlink browser



video

Stratified summary graph



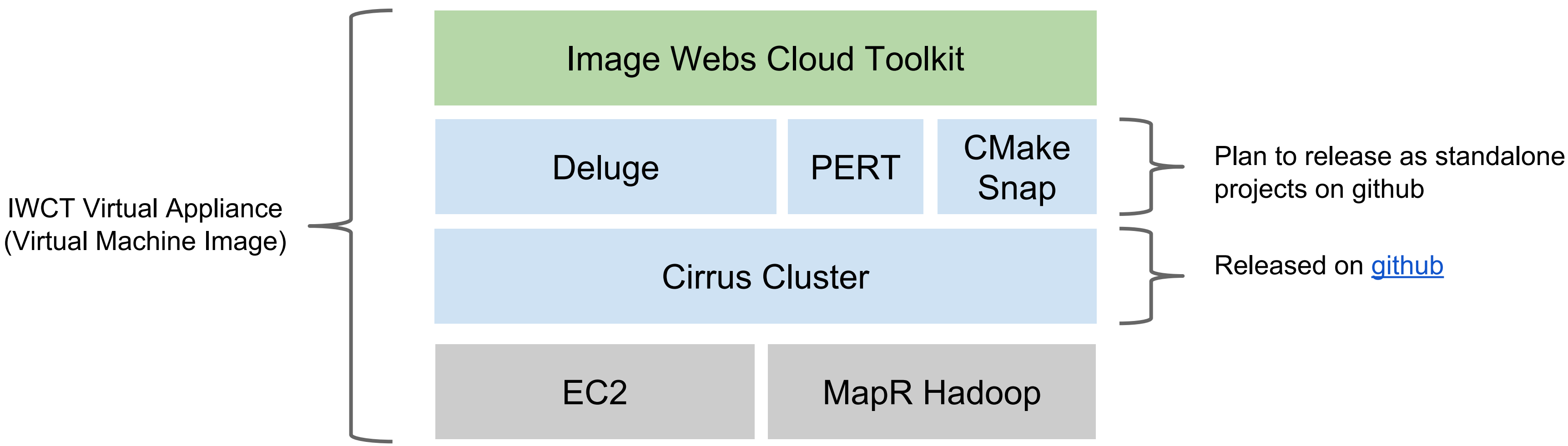
video

Observation: Build environments (cities, buildings) can induces linear structures

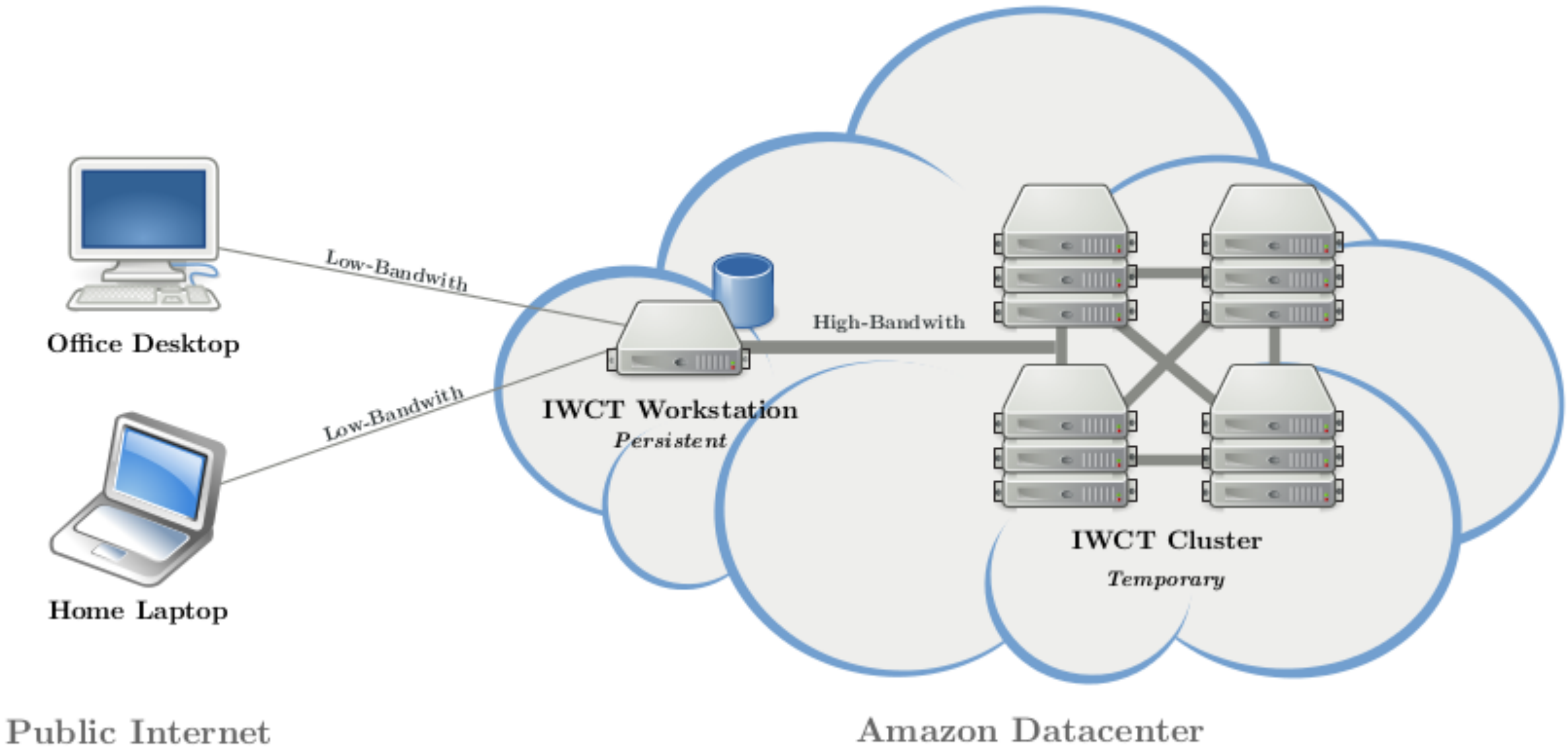
Cloud computing... for researchers

When to use the cloud and some tools to make it easier...

My tools for the cloud...

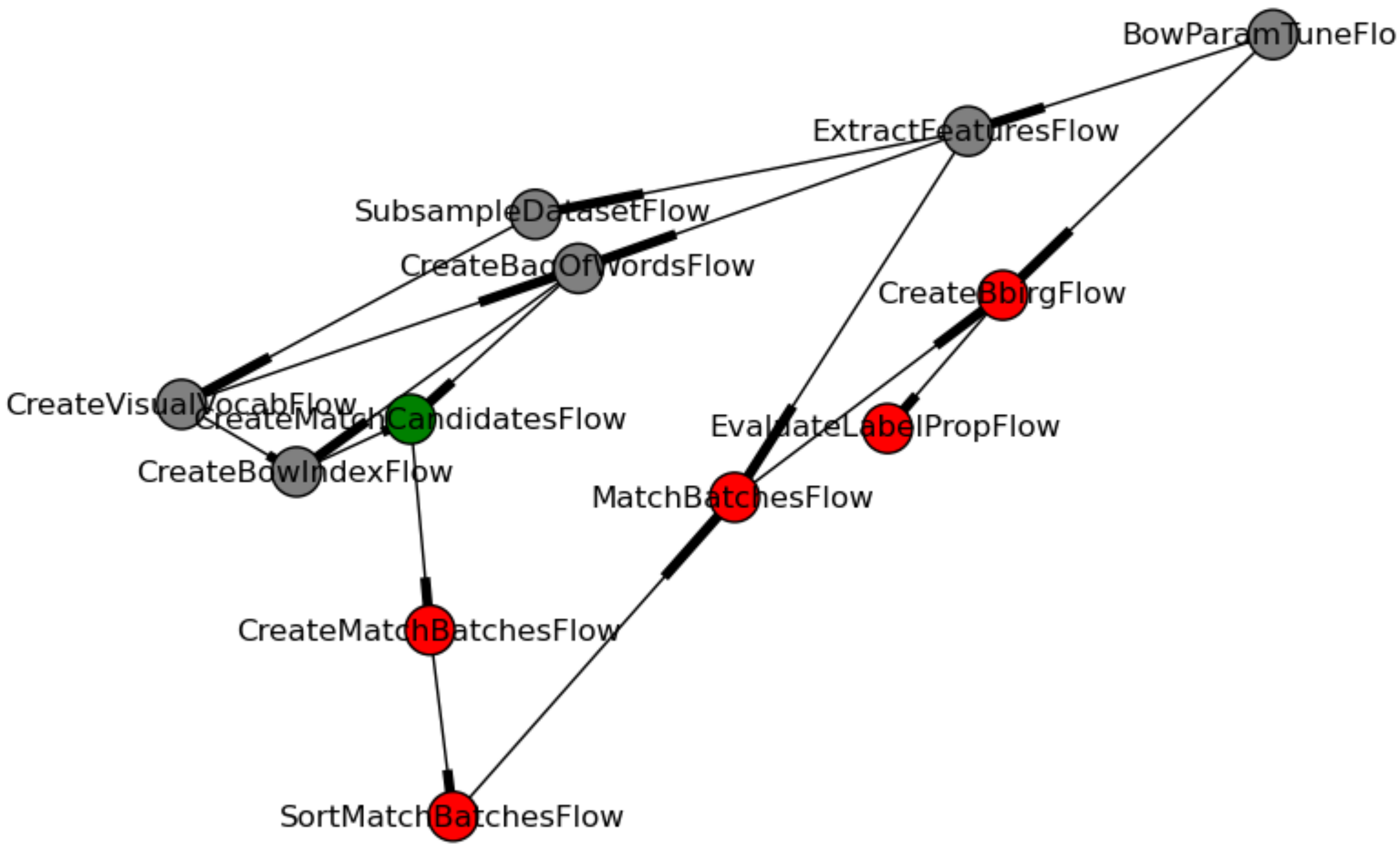


Cirrus Cluster

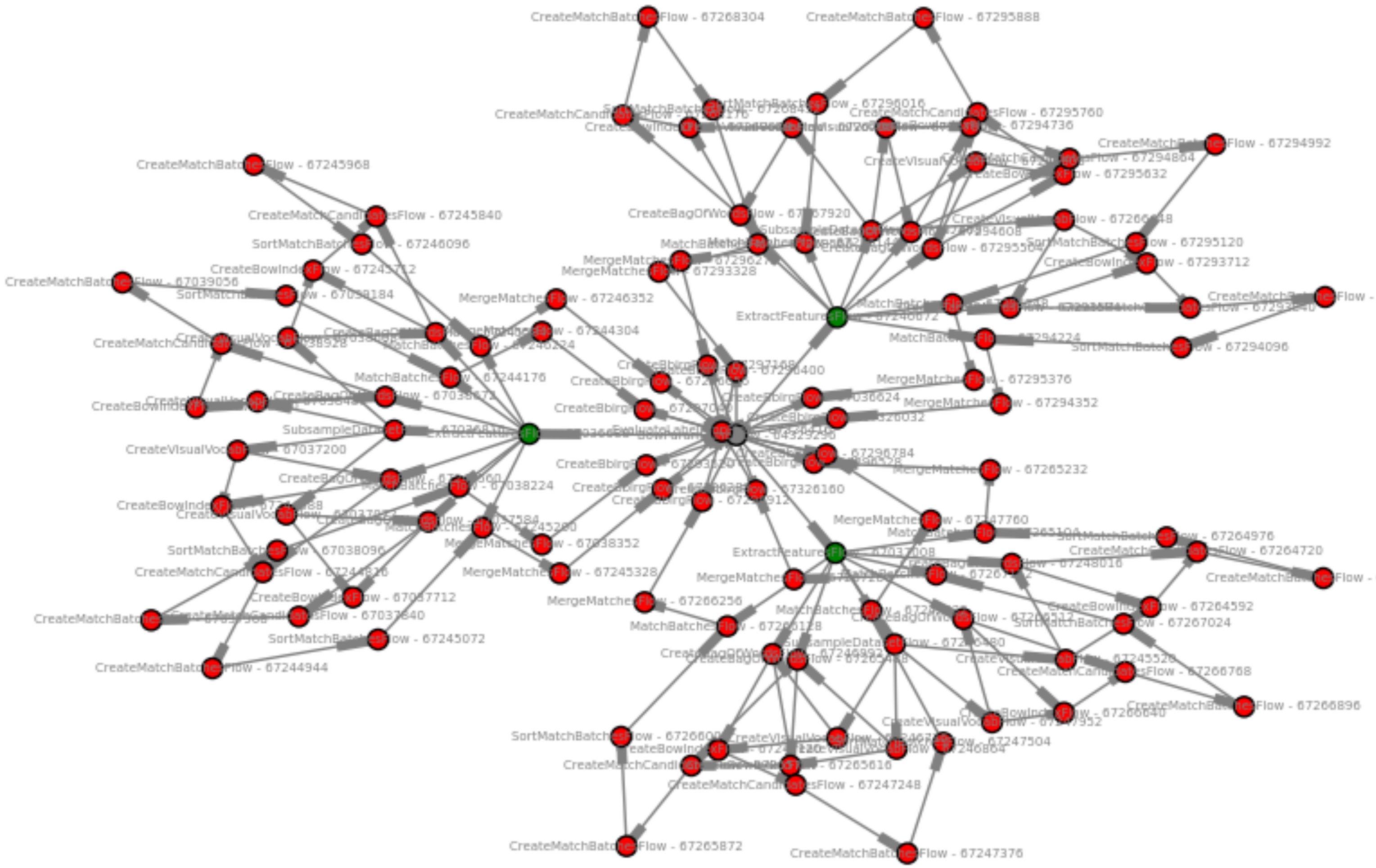


[demo]

Deluge:
Example Map Reduce pipeline



Deluge:
Param Tune Map Reduce pipeline



Should I use the cloud for my research?

Pro

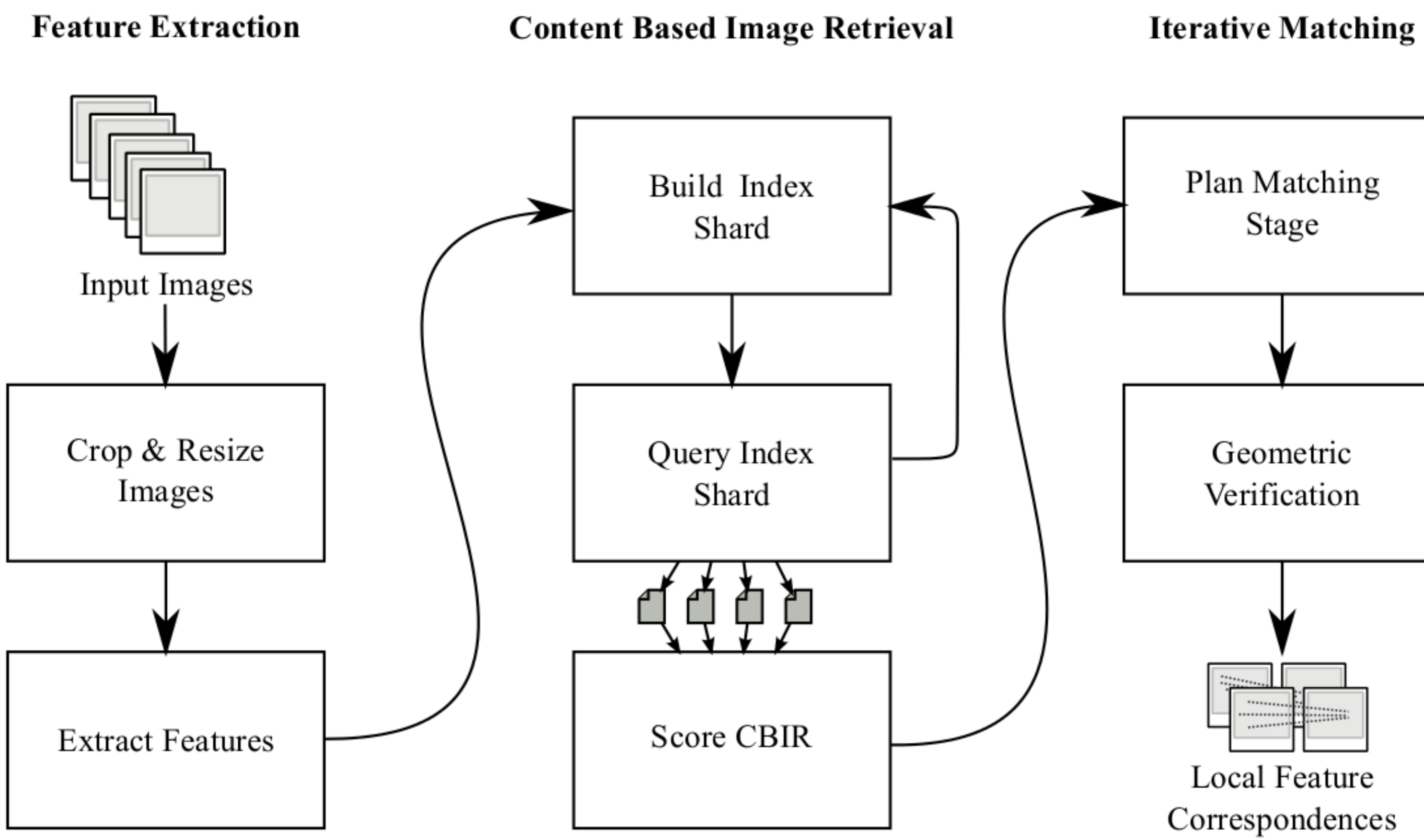
- Others can easily reproduce your results
- Analyze large datasets
- Cheap and getting cheaper!

Con

- Change expense model from hardware to service
- Change workflow model
    - Intermediate results remain in the cloud

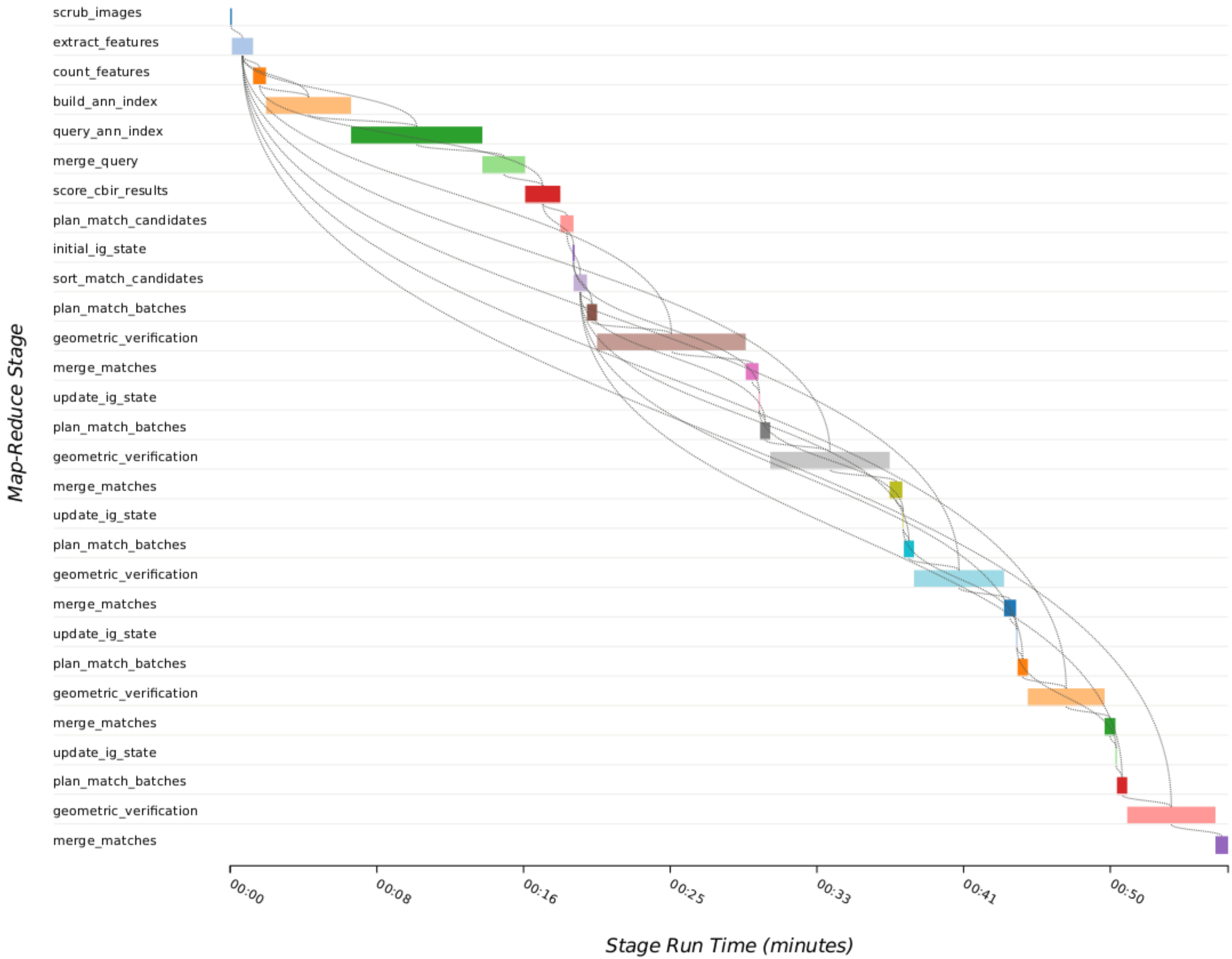
Thanks

Appendix:
IWCT image matching pipeline



Cloud computation is practical!

- Experiment conditions
  - input: 5,000 images
  - processing: 250,000 geometric verifications
  - resources: 6 *c1.xlarge* EC2 instances
- Cost
  - Time: < 1 hour
  - Money: < $3



Map-Reduce Stage
scrub_images
extract_features
count_features
build_ann_index
query_ann_index
merge_query
score_cbir_results
plan_match_candidates
initial_ig_state
sort_match_candidates
plan_match_batches
geometric_verification
merge_matches
update_ig_state
plan_match_batches
geometric_verification
merge_matches
update_ig_state
plan_match_batches
geometric_verification
merge_matches
update_ig_state
plan_match_batches
geometric_verification
merge_matches
update_ig_state
plan_match_batches
geometric_verification
merge_matches
00:00
00:08
00:16
00:25
00:33
00:41
00:50
Stage Run Time (minutes)

Appendix:
Object-recognition results viewers

Tide V2.0 Evaluation - IG

Confusion Matrix

- Drilling down...
  - Success Cases
    - starbucks_logo
    - thinker
    - nasa_spaceshuttle
    - monarch_butterfly
  - Failure Cases
    - vwbug labeled prius
    - violin labeled nasa_spacechuttle
    - thinker labeled unknown

Tide V2.0 Evaluation - IRG

Confusion Matrix

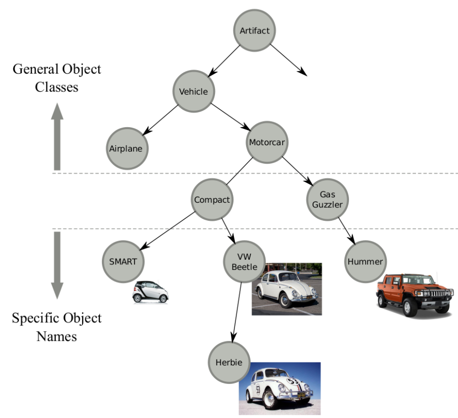
- Drilling down...
  - Success Cases
    - kfc logo
    - monarch_butterfly
    - peacock
    - r2d2
    - tmp
  - Failure Cases
    - kfc logo labeled unknown

Appendix: TIDE dataset

Characterization of TIDE dataset

- TIDE object classes are "fine-grained"
- TIDE is only fine-grained dataset large enough for semi-supervised learning
- TIDE is not artificially 'easy'

TIDE is "fine-grained"

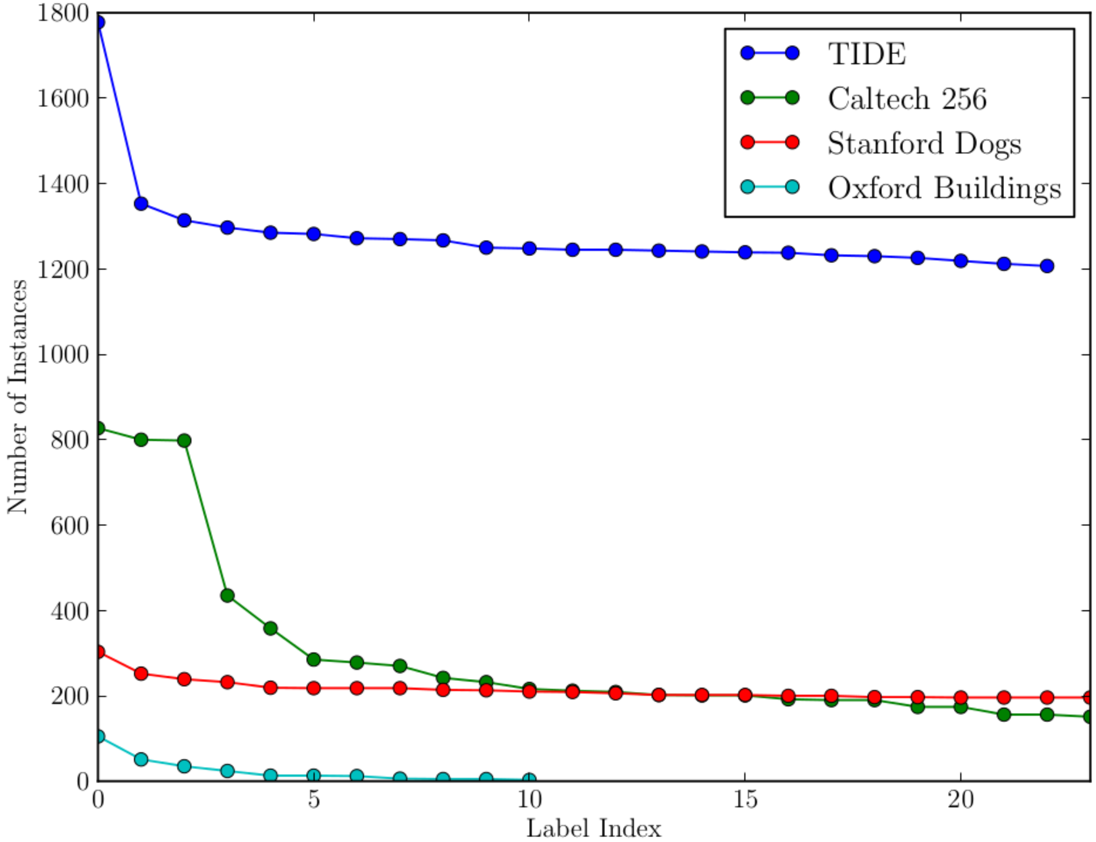


Dataset	Specificity	Example Labels
MSRAMM1	7.79	animal, apple, athlete, baby
NUSWIDE	7.96	airport, animal, beach, bear
Oxford-Buildings	9.18	all-souls, ashmolean, balliol, christ-church
Caltech-256	9.64	ak47, american-flag, backpack, baeball-bat
TIDE	10.95	british-phonebooth, clarinet, clown-fish, csx-locomotive
Stanford-Dogs	15.51	airedale, austrialian-terrier, afghan-hound, african-hunting

Table 7.1: List of datasets ordered by increasing specificity. The specificity score is calculated as the avgerage depth of the label in the WordNet heirarchy.

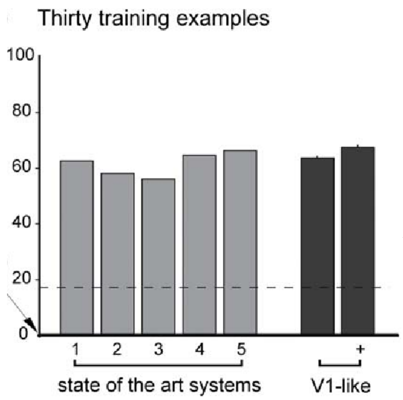
TIDE is a dense sampling

Provides ~ 6x more semi-supervised instances per object



TIDE is not an 'easy' dataset

A state-of-the-art benchmark method finds TIDE-10 "difficult" (40% precision, 10% recall)



Benchmark method
V1-like features + SVM from "Why is Real-World Visual Object Recognition Hard?" By Nicolas Pinto, David D. Cox and James J. DiCarlo (2008)

- As good as far more complex state-of-the-art methods
- Quality source code available

Appendix:
Feature Matching Tips and Tricks

RootSIFT *

- Trivial transform of standard SIFT descriptor
- Significantly improves matching
- Why it matters
  - Accurate alternatives to L2 distance preclude accelerated search techniques
  - RootSIFT distance = L2 on normalized descriptors... Use existing search acceleration tools!

* Arandjelovic, Relja, and Andrew Zisserman. "Three things everyone should know to improve object retrieval." *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*. IEEE, 2012. (PDF)

RootSIFT Example

Standard Sift

(no model found)

RootSift

(model found)

nfa=-12.608469 num_matches=17 precision=10.851171

Full-Representation CBIR is better *

- Motivation for BOW was to compress index to fit in RAM of a single machine
  - At large scale... must span many machines anyway
- Cost of compression is quantization noise
  - Much effort spent trying to recover lost performance

* Aly, Mohamed, Mario Munich, and Pietro Perona. "Indexing in large scale image collections: Scaling properties and benchmark." *Applications of Computer Vision (WACV), 2011 IEEE Workshop on*. IEEE, 2011. (PDF)

AC-RANSAC is major improvement

- Standard RANSAC is brittle
  - Performance sensitive to a set of coupled parameters
    - No fixed set of parameters suitable for range of object classes
- AC-Ransac *
  - Uses a-contrario principle to select suitable RANSAC parameters for each candidate pair
  - One (interpretable) parameter to rule them all!
    - Expected Number of False Alarms (NFA)

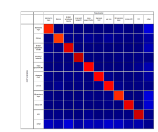
* Rabin, Julien, et al. "MAC-RANSAC: a robust algorithm for the recognition of multiple objects." *Proceedings of 3DPTV 2010* (2010). (PDF)

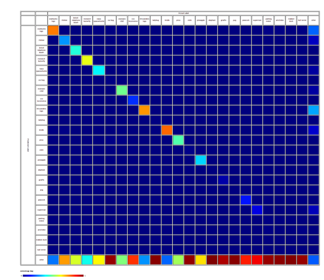
Appendix:
What was wrong with initial design?

Problem: Matching pipeline didn't scale

10 objects : 20,000 images
2000 positive instances
2000 relevant instances
16000 distractors

23 objects : 313,942 images
4600 positive instances
23000 relevant instances
208658 distractors





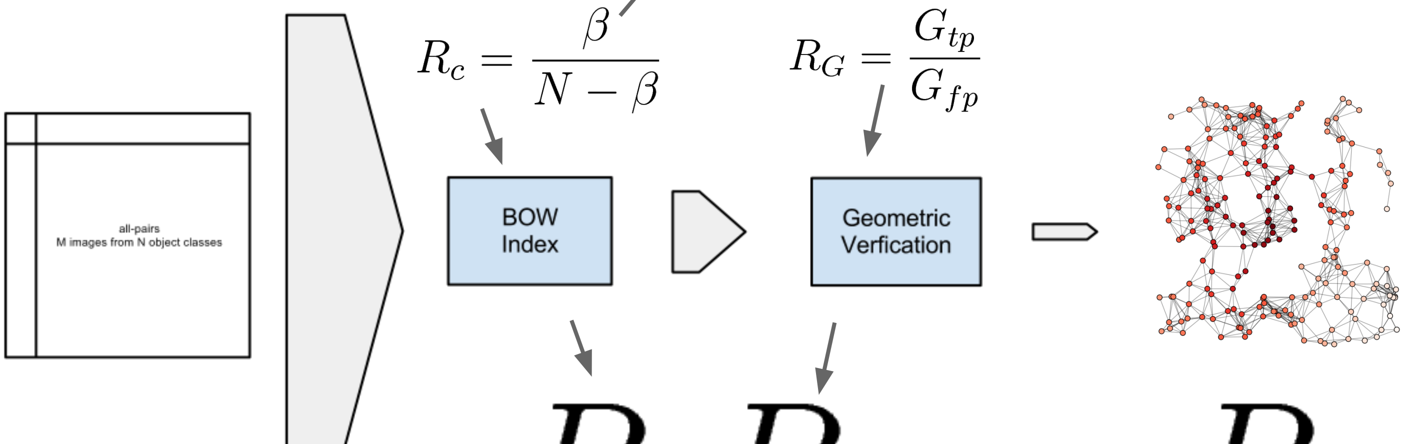
Model of correct / incorrect edges through matching pipeline...

Given an image dataset sampled uniformly from N object classes...

CBIR precision improvement over "random chance"

$$R_c = \frac{\beta}{N - \beta}$$

$$R_G = \frac{G_{tp}}{G_{fp}}$$



$$R_c R_G = R_e$$

Ratio of
Correct / Incorrect
Edges

Model of correct / incorrect edges through the image web pipeline...

Given an image dataset sampled uniformly from N object classes...

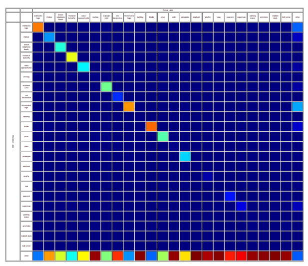
To preserve performance while scaling...

- Increase CBIR performance
- Increase Geometric Verification performance

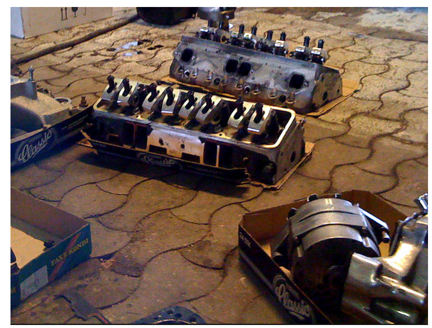
Edges

Improving image matching pipeline...

Doesn't work well on TIDE V2.0



Take matching engine apart



Swap out components

Works well on TIDE V2.0

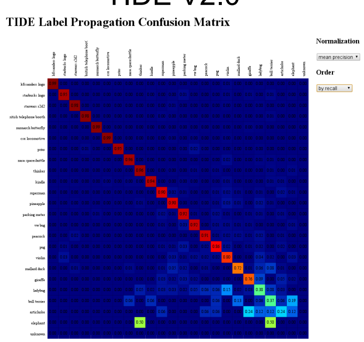


Image matching pipelines upgrades...

TIDE Dataset → CBIR → Geometric Verification → Evaluation

"Instagram" border removal
- ~~Border detectors~~
- 5% crop

Compute Infrastructure

Custom built tools
- Cirrus - Launch cluster
- Deluge - Hadoop flows
- CMakeSNAP - Simpl

MAPR
- ~~v 2.0.0~~
- v 2.1.1

Cluster Geometry
- ~~High-CPU instance types~~
- New HPC instance types

ANN Search
- ~~FLANN + single thread~~
- ~~FLANN + Intel TBB~~
- FLANN + Boost Thread Pool

Feature matching criteria
- ~~Fixed distance~~
- ~~Lowe's Ratio Test~~
- ~~AC Background DB~~
- k-NN

Doc Scoring
- ~~Count~~
- Adaptive

Score Normalization
- ~~None~~
- NRC

Geometric Re-Ranking
- WGC (S)
- WGC (S+T)
- Smoothness

RANSAC
- ~~Standard RANSAC~~
- ~~Mesh-RANSAC~~
- AC-RANSAC

SIFT Distance Metrics
- ~~Circular Earth Mover Distance~~
- RootSIFT

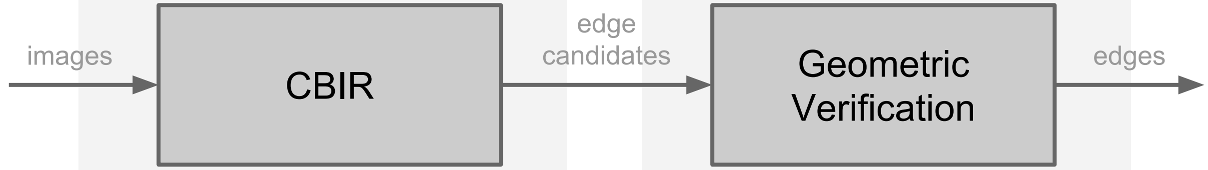
Construction Strategies
- ~~One shot Top N~~
- ~~Iterative by CBIR score~~
- Iterative by CBIR rank
- Motif expansion

CBIR Stage Eval
- ~~Precision alone~~
- mAP on TIDE
- Confusion matrix

GeoVerifcation Stage Eval
- All-pairs graph on single TIDE class

Graph eval tools
- ~~Edge count stats~~
- Edge confusion matrix viewer
- Image graph viewer
- Label propagation simulations

Main Improvements



	CBIR	Geometric Verification
Method	**Bag-Of-Words Representation**	**RANSAC**
Problem	Quality* of edge candidates generated generated decreases rapidly with dataset complexity... *(absolute number and proportion that are within-class)	No fixed set of parameters that achieves good precision and recall across object classes and dataset size.

Main Improvements