Joint Inference in Image Databases via Dense Correspondence

Michael Rubinstein
MIT CSAIL
(while interning at Microsoft Research)

My work

- Throughout the year (and my PhD thesis): **Temporal Video Analysis and Visualization**

Pulse signal amplified

Breathing motions amplified

- This short talk: my work during the summers (MSR 2011, 2012)
  - Inference in large, weakly-annotated image databases

Videos vs. Image Datasets

- Goal: we want to infer properties of pixels/regions
  - Semantics, layers, geometry (depth), motion, …
- Recent advances allow us to **treat a set of images like videos**!
  - Correspondence between **adjacent frames in videos**: optical flow, layer models, tracking, …
  - Correspondence between **similar images in databases**: Feature Matching, graph matching, Spatial Pyramid Matching (SPM), SIFT flow, …

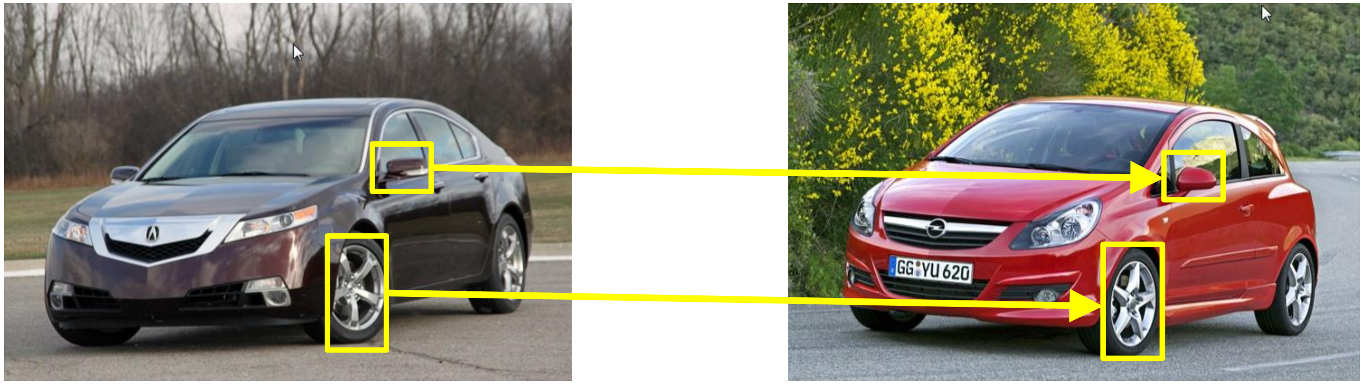


Image Correspondence is Challenging…

Query Best match

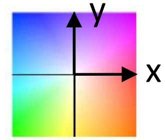
Multiple objects; no global transform

Changing perspective, occlusions

Intra-class variation

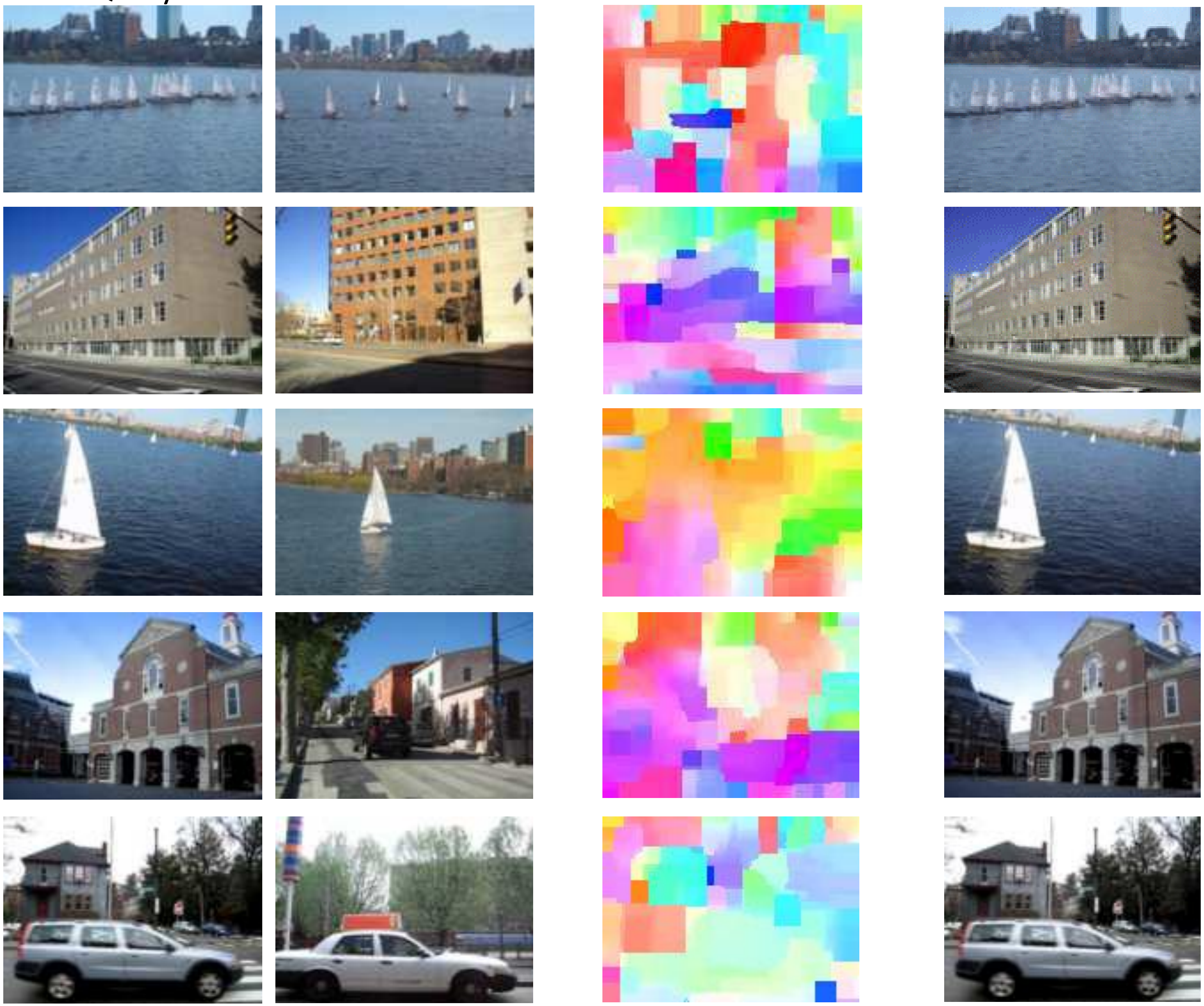
Background clutter

…but Good Solutions Exist

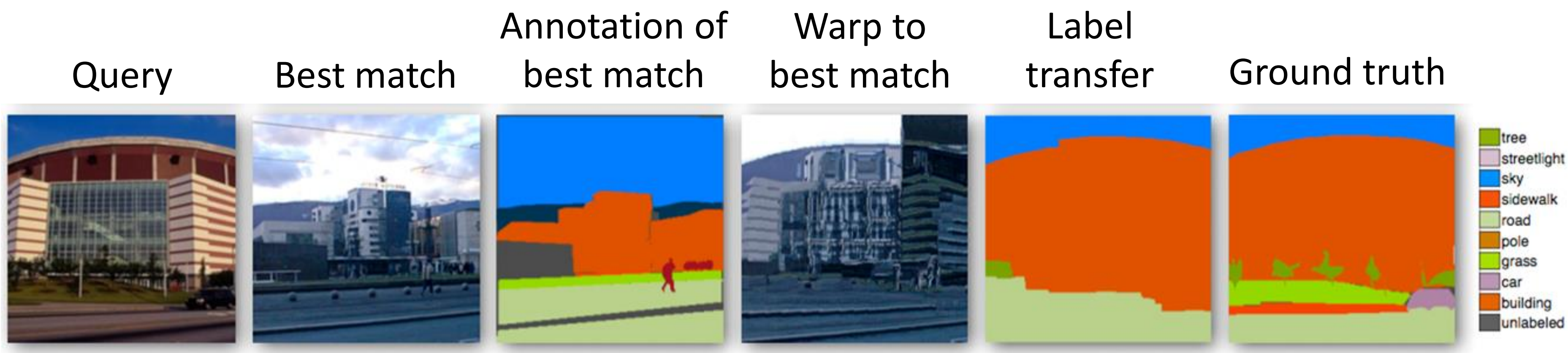


SIFT Flow [Liu et al. TPAMI 2011]

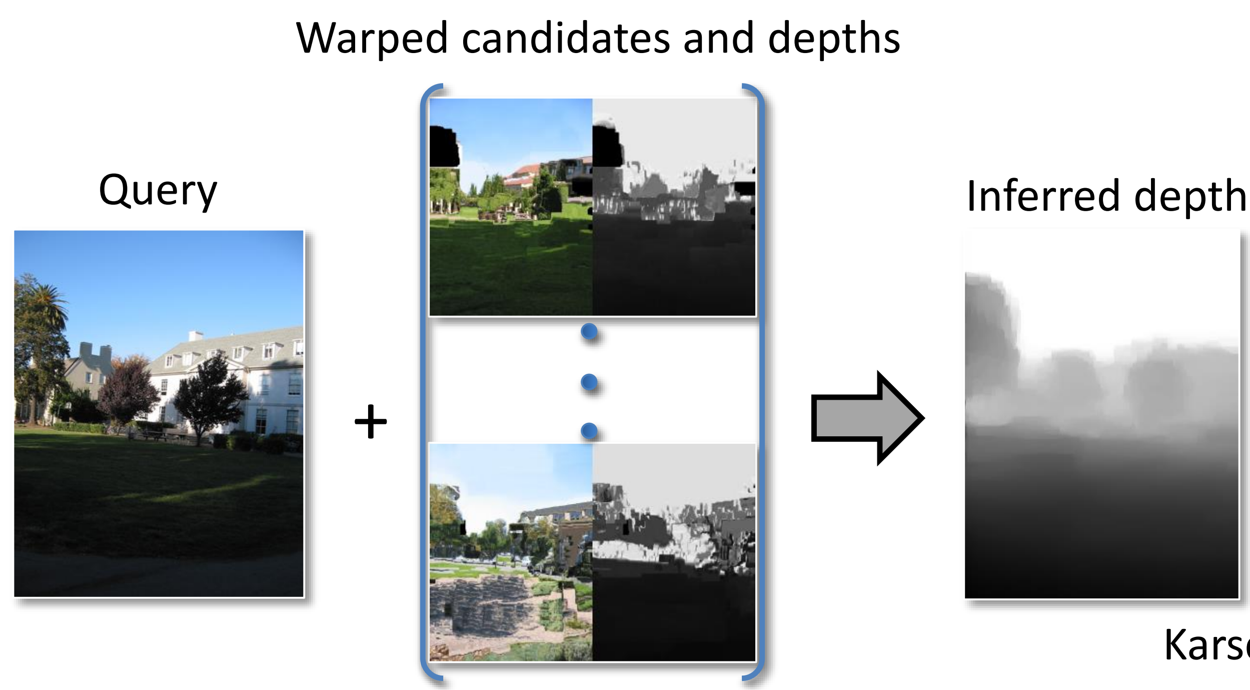
Query Best match



Correspondence-driven Approaches to Computer Vision

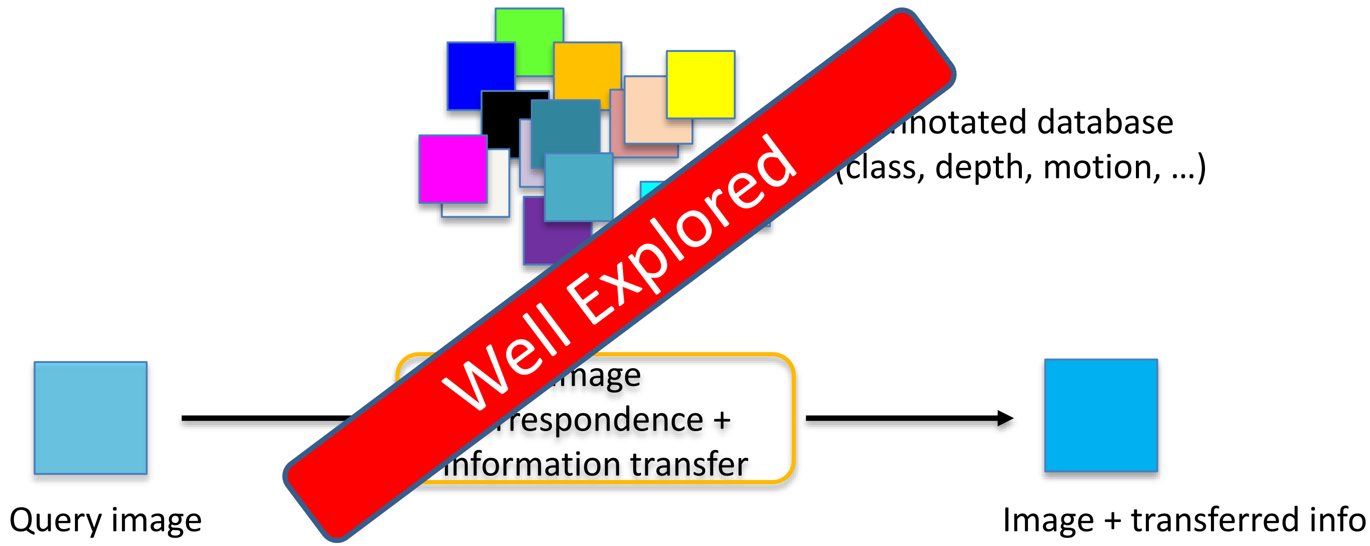


Liu et al. TPAMI’11



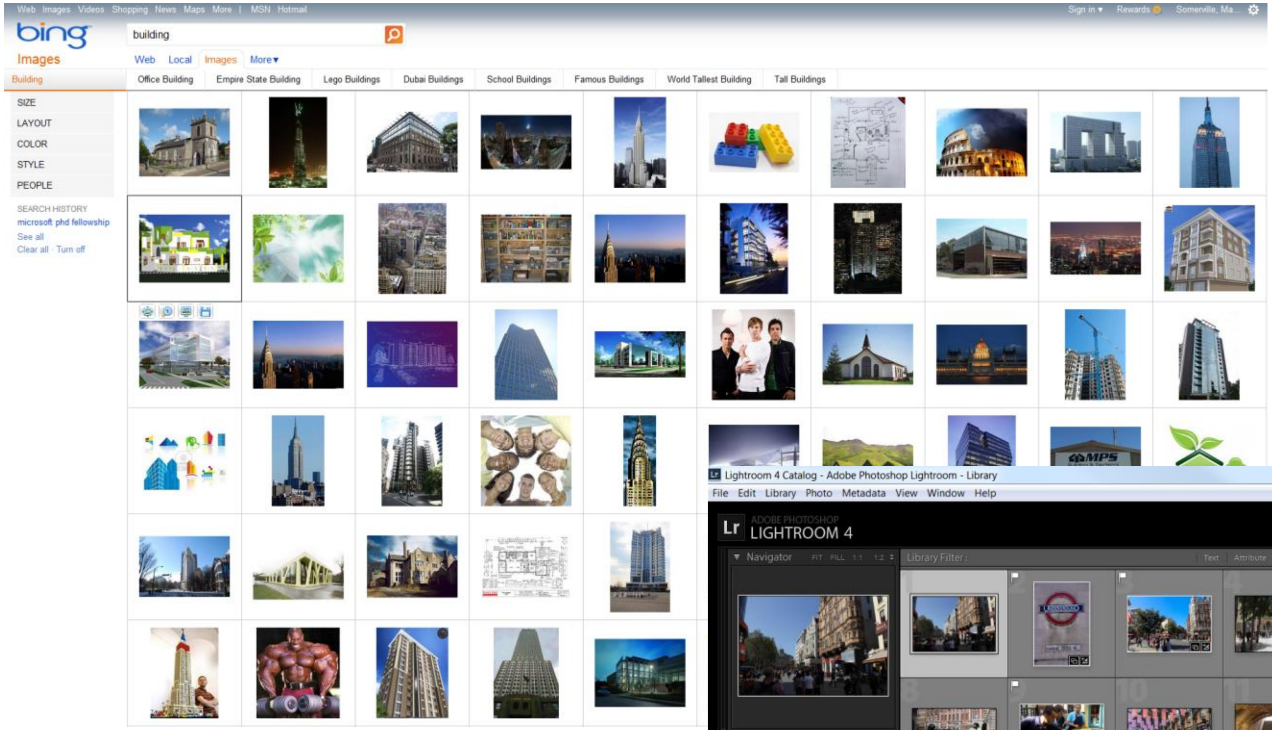
Karsch et al. ECCV’12

How to *densely* label new images?



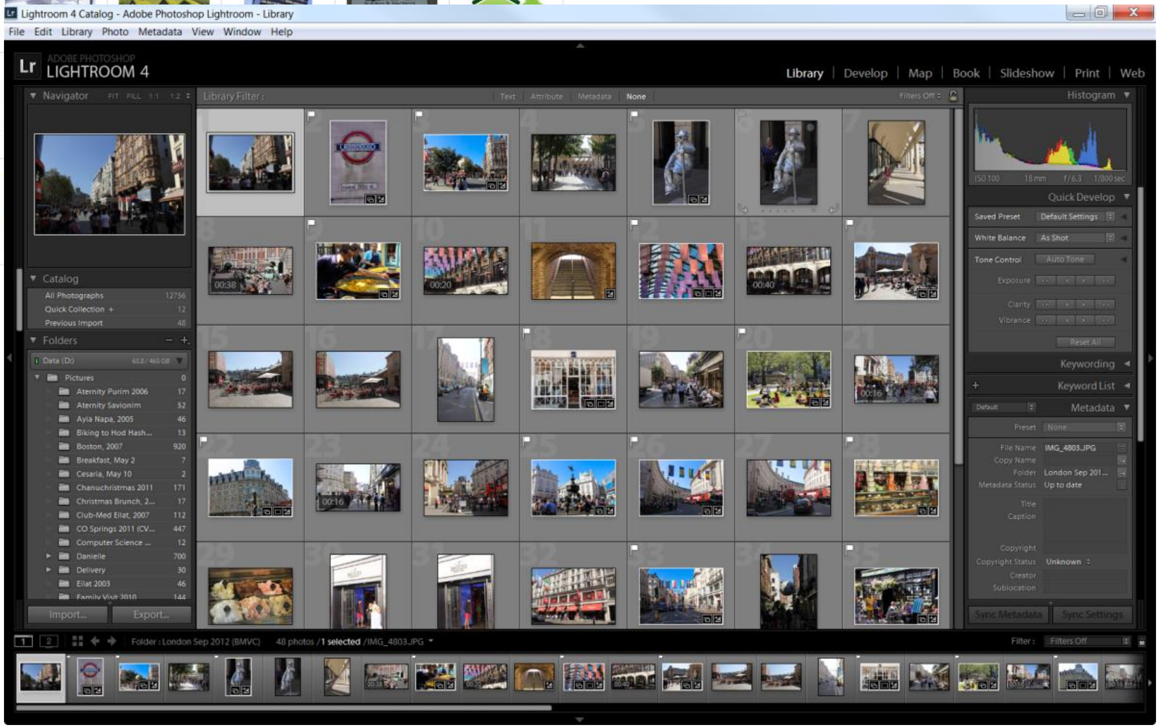
Big Visual Data

Pixel labels usually unavailable!

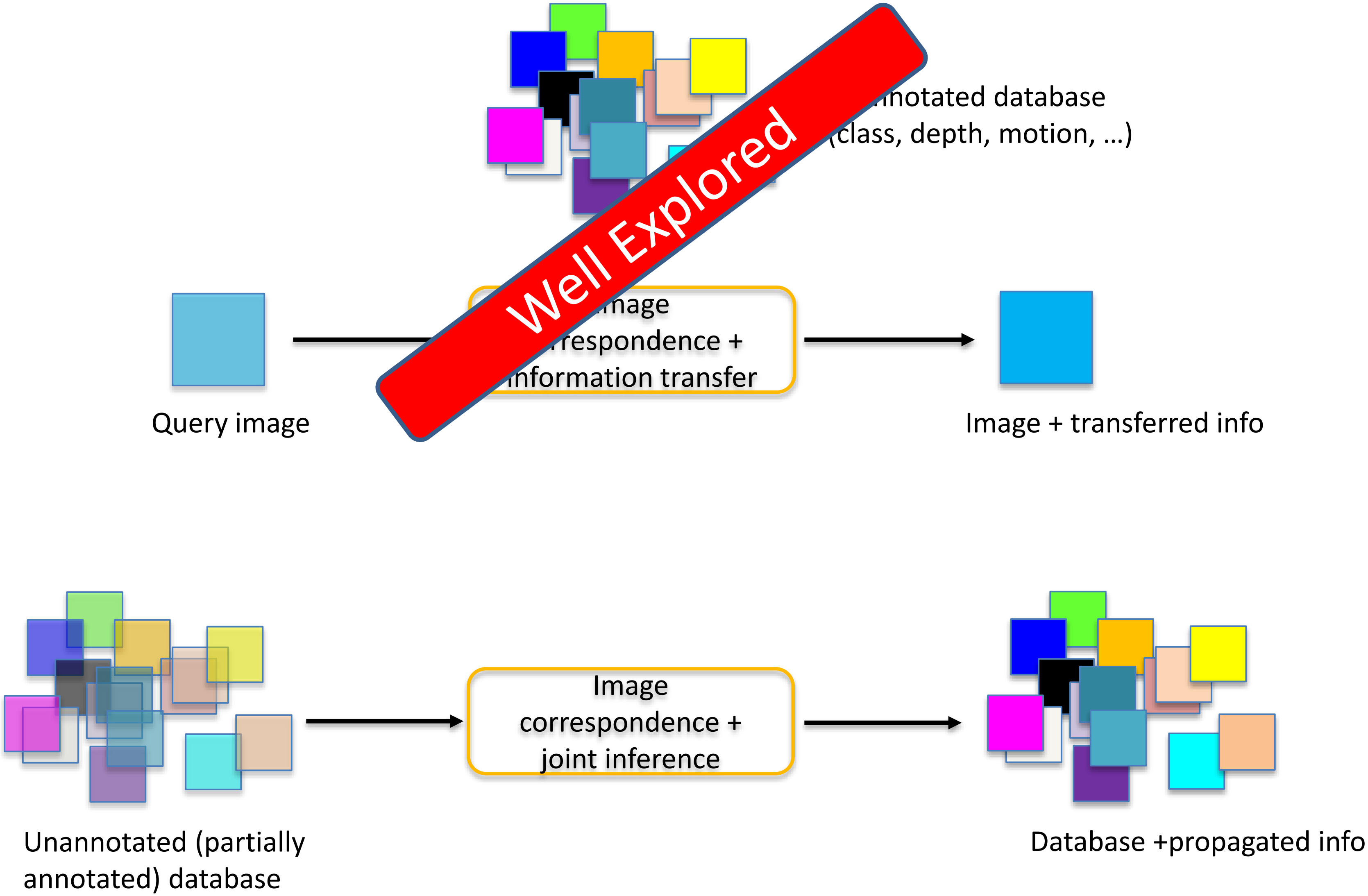


Internet

Photo collections



How to *densely* label new images?



Joint Inference for Image Databases

- <u>Weakly supervised</u>

Annotation Propagation in Large Image Databases via Dense Image Correspondence (ECCV 2012)

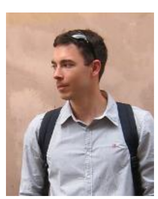
With Ce Liu, William T. Freeman

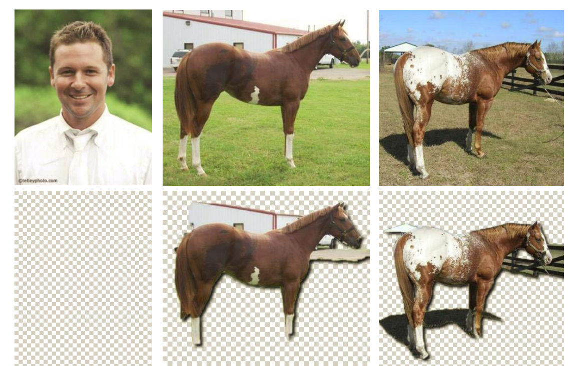

- <u>Unsupervised</u>

Unsupervised Joint **Object Discovery and Segmentation** in Internet Images (CVPR 2013)

With Ce Liu, Armand Joulin, Johannes Kopf

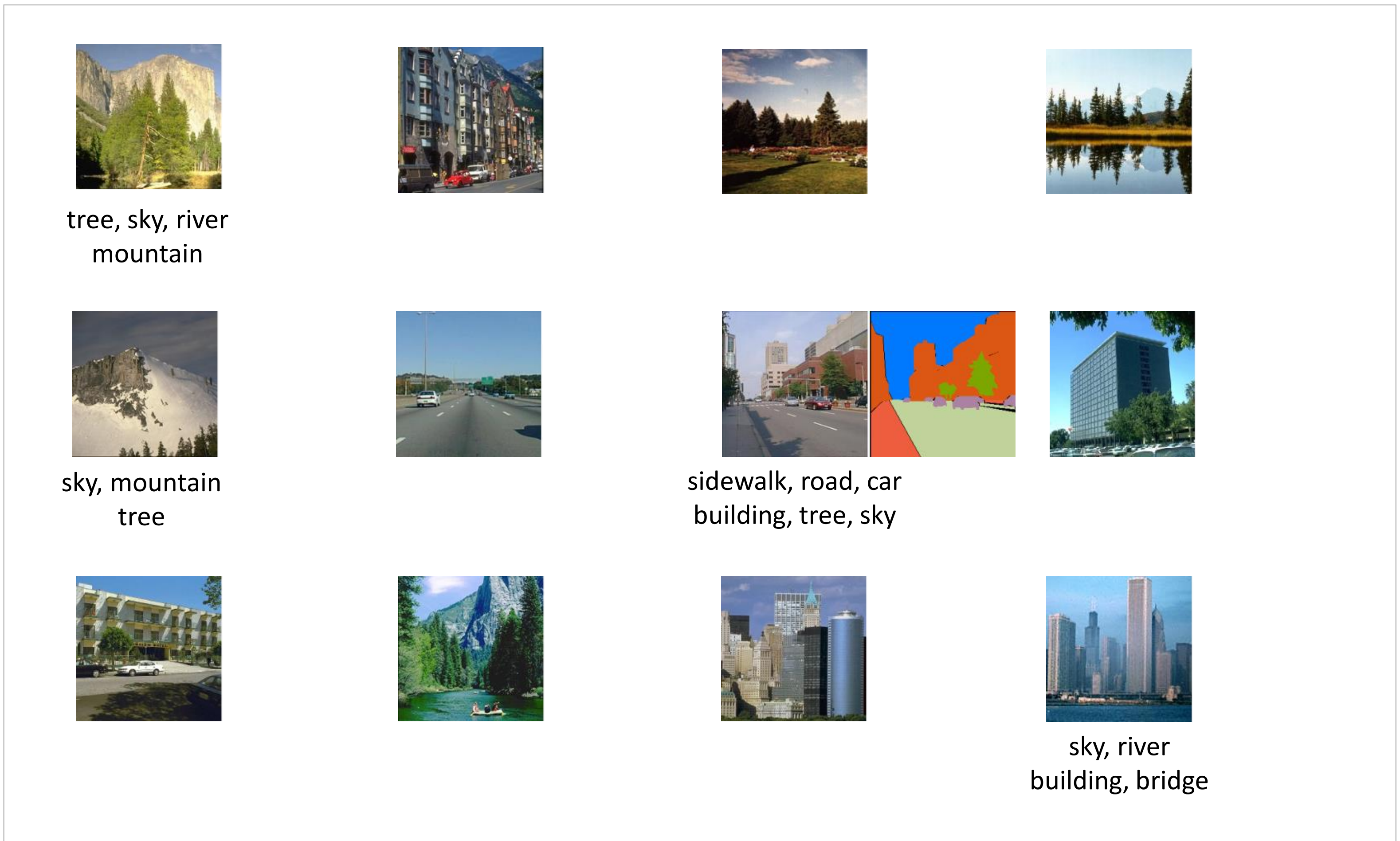






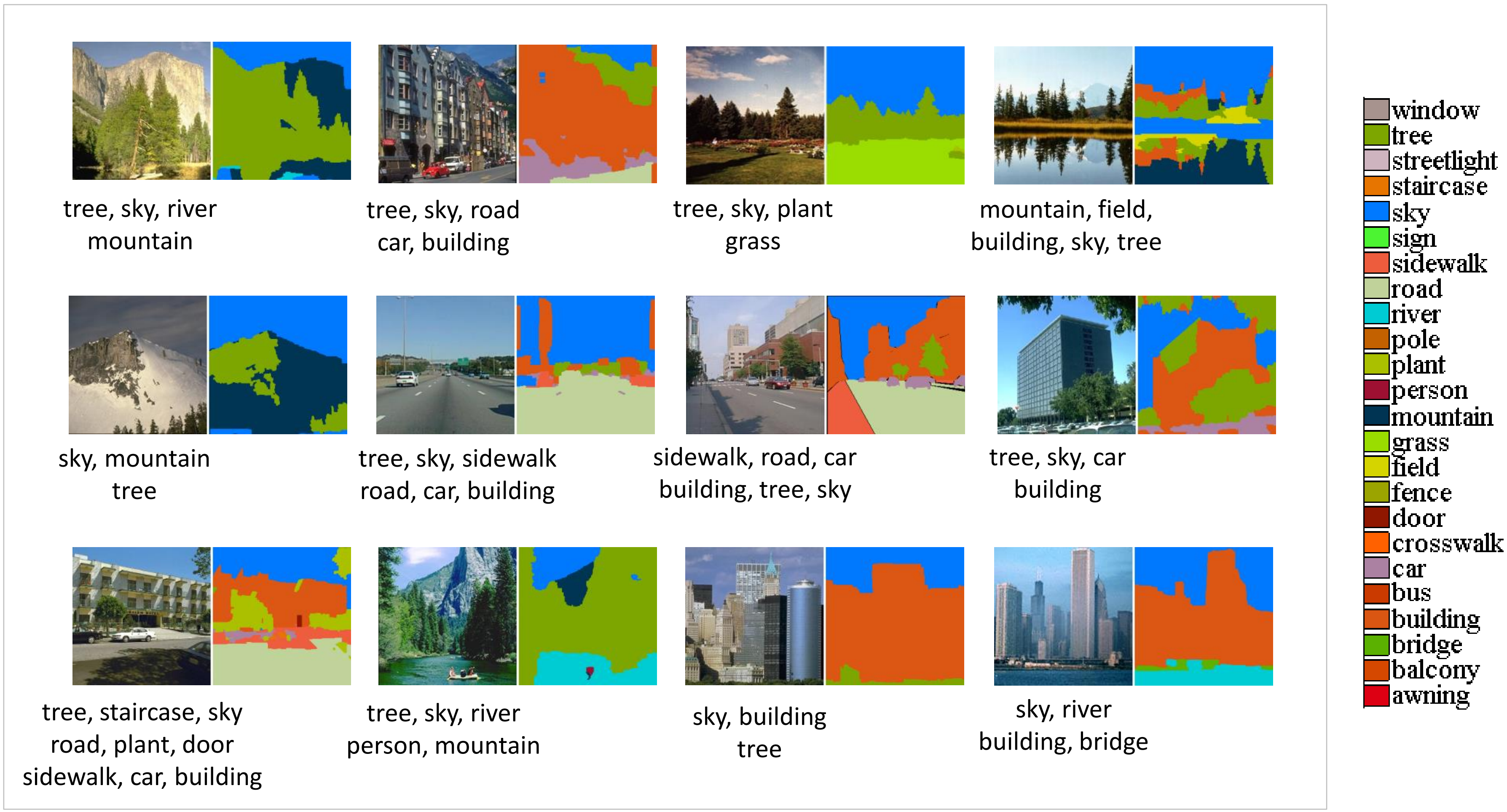
Annotation Propagation

<u>Input:</u> A large database of images where only some are tagged and very few (possibly none) are densely labeled



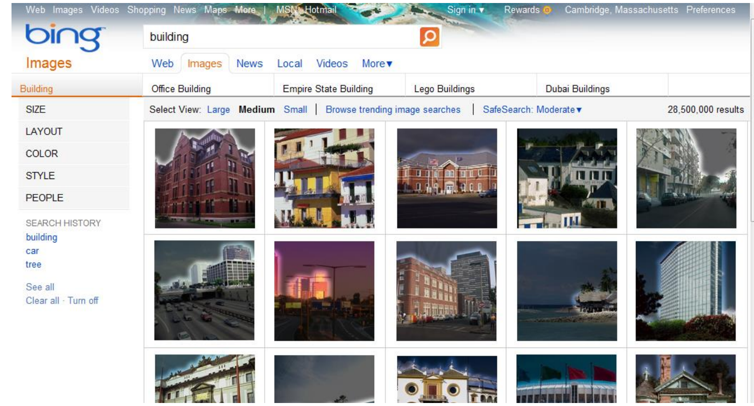
Annotation Propagation

<u>Output:</u> The same database with all the pixels labeled and all the images tagged



Dense pixel/region labeling is important

- Enhanced image search



- Constructing training sets for detectors/classifiers

PASCAL 2012

- Image editing
  - User edit propagation

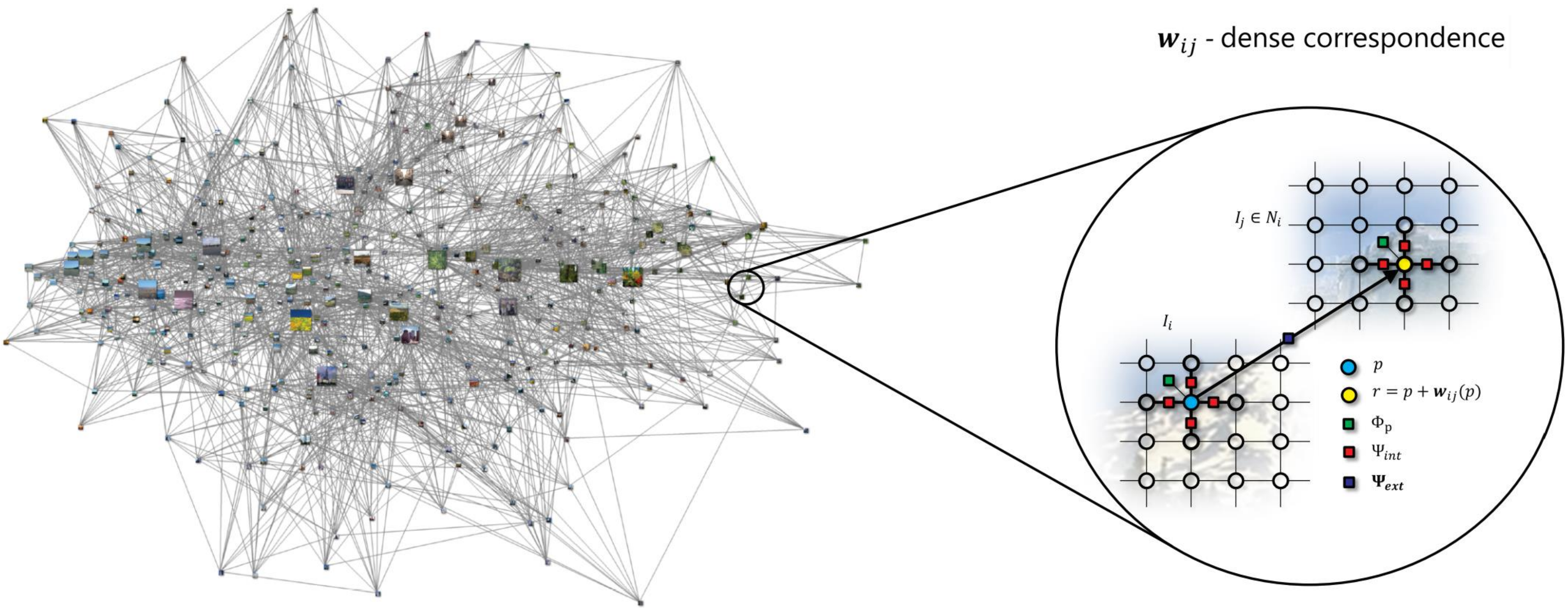
HaCohen et al. 2013

Pixel-wise image graph

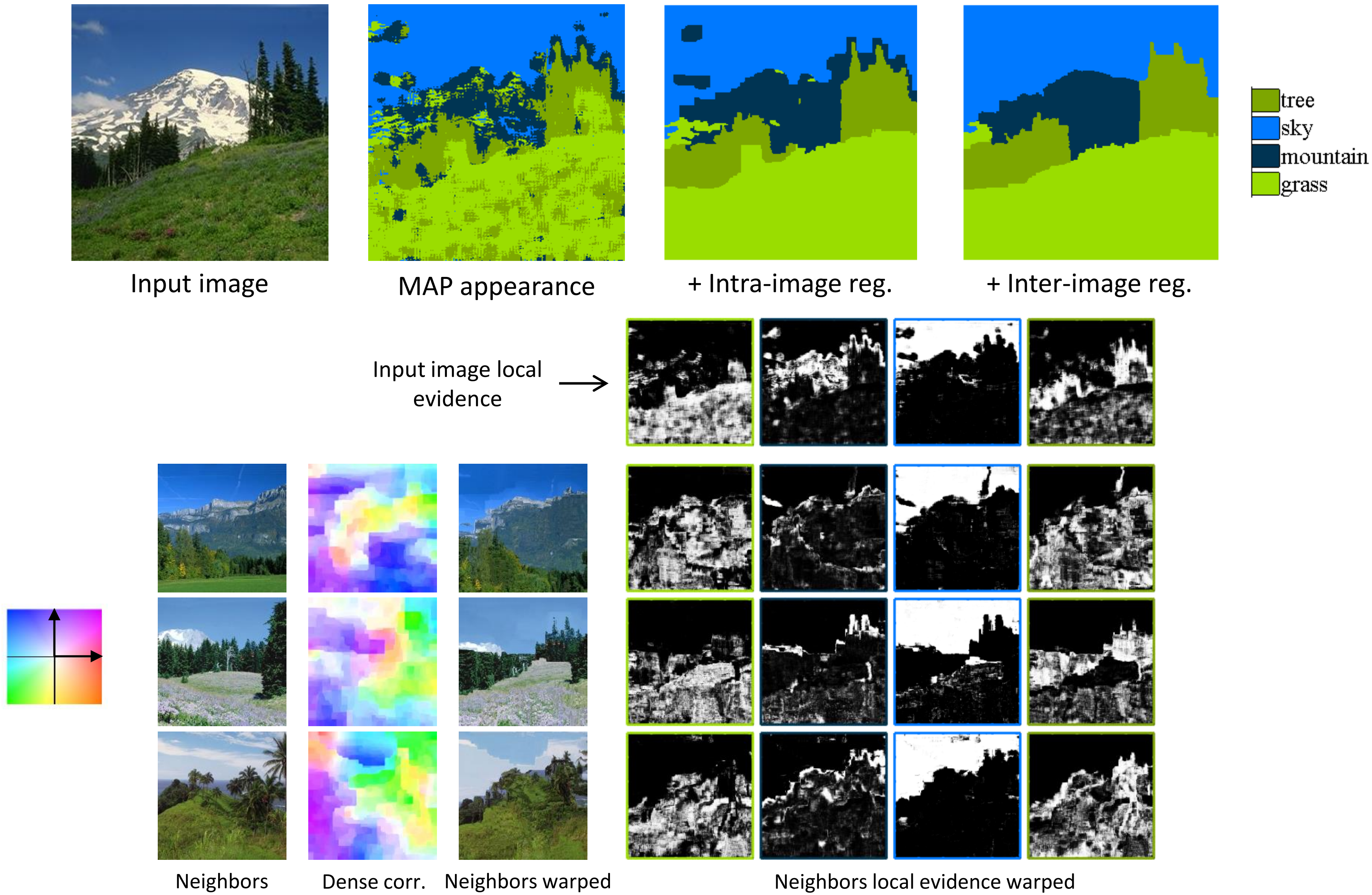
$P(\text{word} \mid I(\boldsymbol{p}))$ – using machine learning

$$E(\boldsymbol{C}) = \sum_{i=1}^{N}\sum_{\boldsymbol{p}\in\Lambda_i}\left[\underbrace{\Phi_p(c_i(\boldsymbol{p}))}_{\text{Local evidence}} + \sum_{\boldsymbol{q}\in N_{\boldsymbol{p}}}\underbrace{\Psi_{int}(c_i(\boldsymbol{p}), c_i(\boldsymbol{q}))}_{\text{Intra-image regularization}} + \sum_{j\in N_i}\underbrace{\Psi_{ext}\left(c_i(\boldsymbol{p}), c_j\left(\boldsymbol{p}+\boldsymbol{w}_{ij}(\boldsymbol{p})\right)\right)}_{\text{Inter-image regularization}}\right]$$

$\boldsymbol{w}_{ij}$ - dense correspondence



Inference Results



Optimization

1. Initialize from given tags and labels
2. Repeat until convergence
    - 2.1. Update appearance model parameters $\boldsymbol{\Theta}$ and compute local evidences $P(c_i(\mathbf{p}); \boldsymbol{\Theta})$
    - 2.2. For each image, repeat until convergence
        - 2.2.1. Intra-image message passing
        - 2.2.2. Update color model $\mathbf{h}^c_{i,l}$
    - 2.3. Inter-image message passing
    - 2.4. Compute MAP label estimate $C^* = \arg\min_C E(C)$
    - 2.5. Update spatial prior $\mathbf{h}^s_l$ and co-occurrence matrix $\mathbf{h}^o$ from C^*
3. Output C^*

Appearance Modeling

Propagation

- Coordinate descent, iterating between estimating the appearance model (learning) and tag propagation (inference)

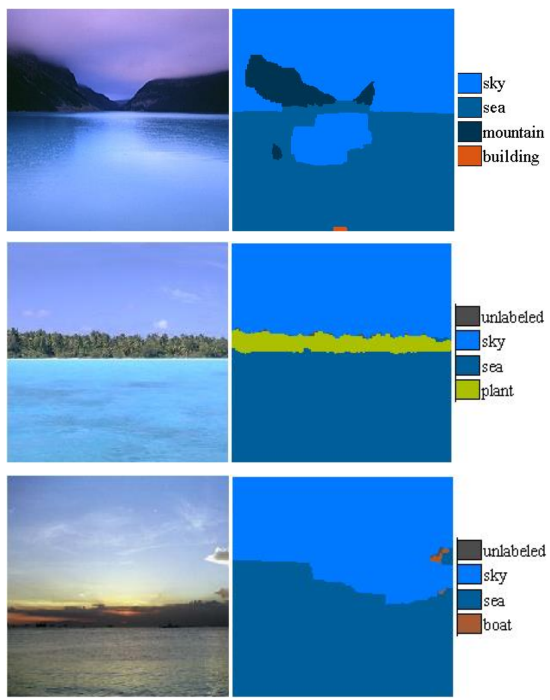
- Lots of engineering, but nothing revolutionary
    - Partition message passing into intra- and inter-image updates
    - Intra-image message passing on separate cores
    - Parallel inter-image message passing

From stronger local evidence to weaker local evidence

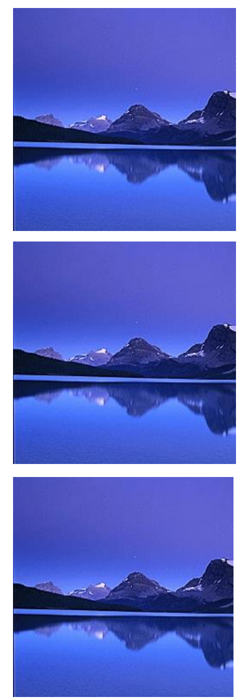
Input image

Local evidence + intra-image reg.

+ Inter-image reg.

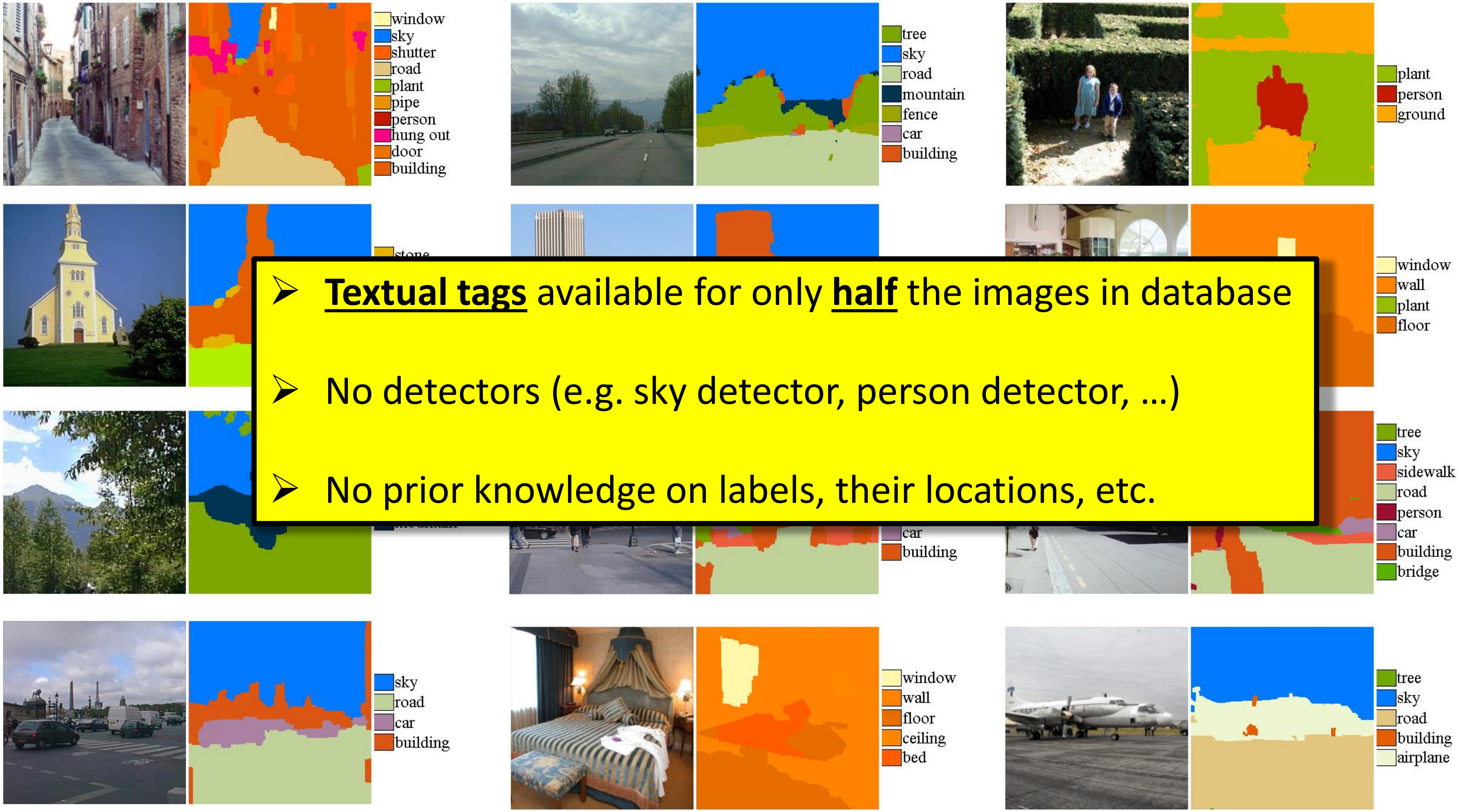


neighbors



warp

Results on SUN Dataset



- **<u>Textual tags</u>** available for only **<u>half</u>** the images in database
- No detectors (e.g. sky detector, person detector, …)
- No prior knowledge on labels, their locations, etc.

SUN dataset [Xiao et al. 2010] - 9556 images, 522 labels

Joint Inference for Image Databases

- Weakly supervised

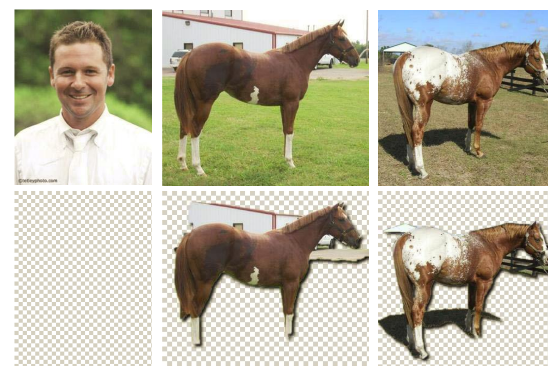
Annotation Propagation in Large Image Databases via Dense Image Correspondence (ECCV 2012)

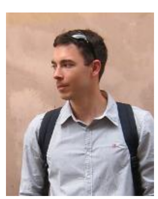
With Ce Liu, William T. Freeman

- Unsupervised

Unsupervised Joint **Object Discovery and Segmentation** in Internet Images (CVPR 2013)

With Ce Liu, Armand Joulin, Johannes Kopf





Object discovery and Co-segmentation

- **Input**: A set of images containing some “common object”
- **Output**: Every pixel in the dataset marked as belonging or not belonging to the “common object”
- **No additional information on the images or the object class**

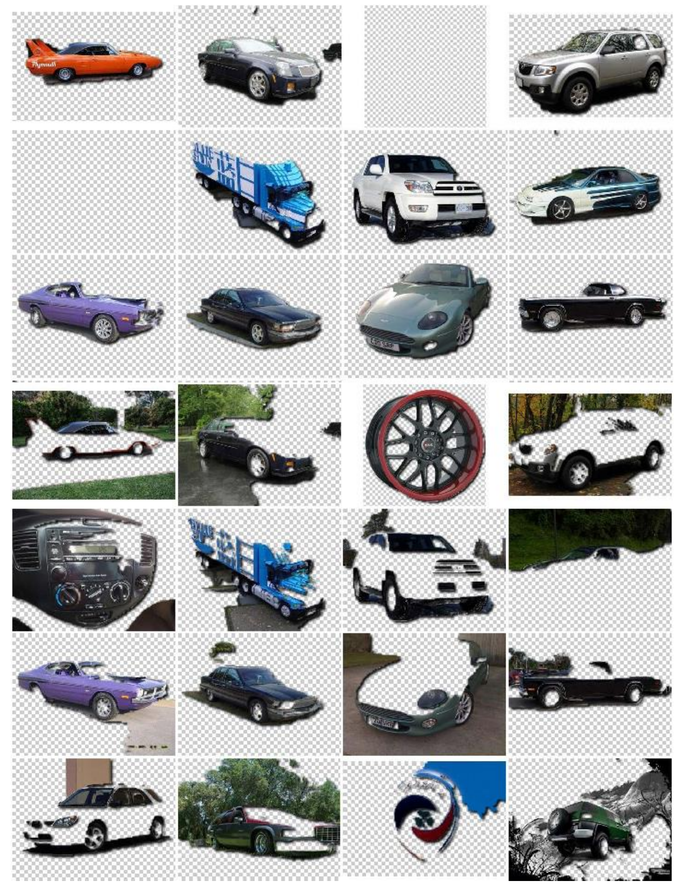
Object discovery and Co-segmentation

“Car”

Image search

Images downloaded from the Internet

Object discovery and segmentation

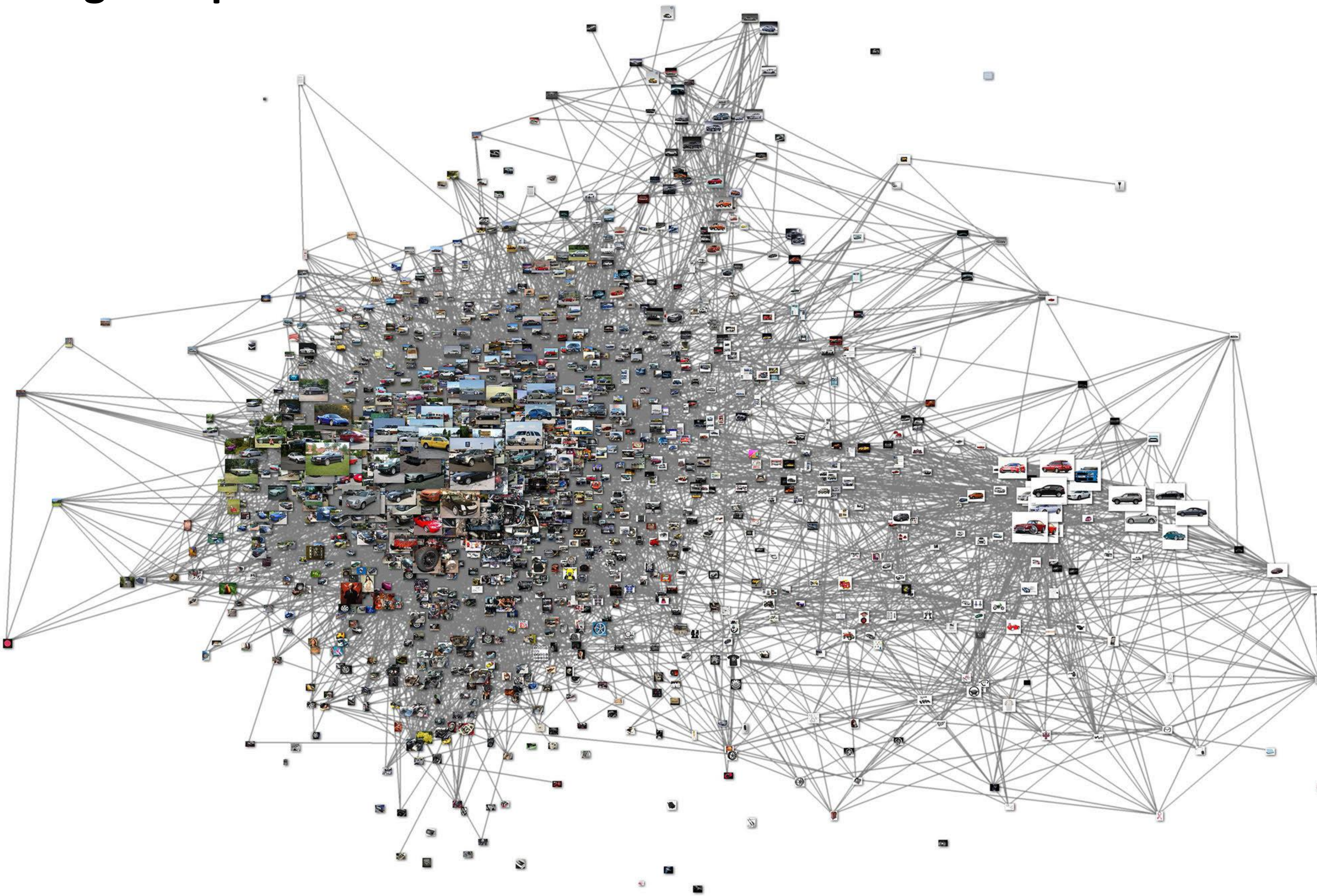


State of the art co-segmentation [Joulin et al. CVPR 2012]

Benchmark “plane” Dataset (MSRC)

Real-world “plane” Dataset (Internet Search)

Image Graph



Basic Idea

- Pixels (features) belonging to the common object should be:

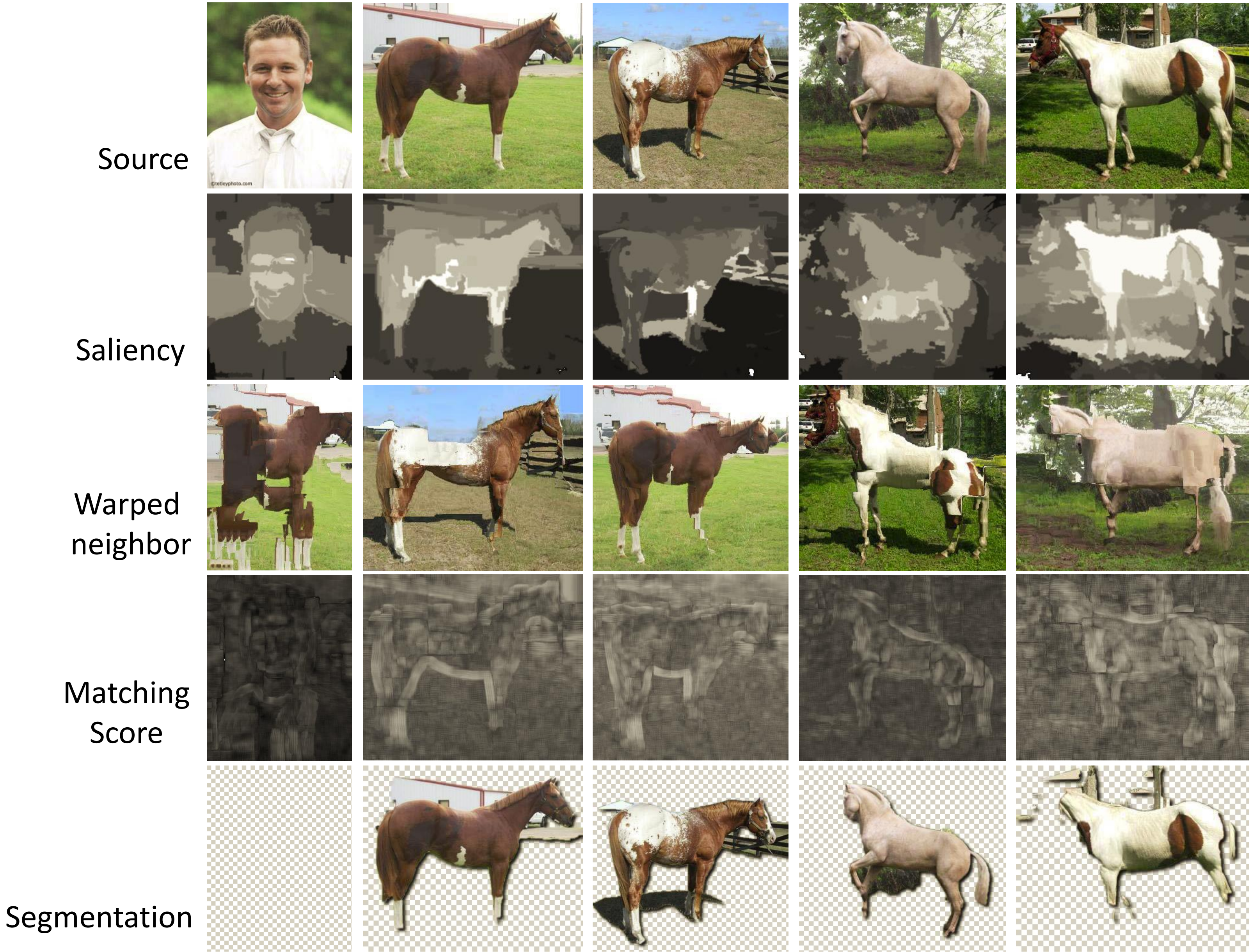
1. ***Salient*** - Dissimilar to other pixels (features) in <u>their image</u>

   Captured by image ***saliency measures***

2. ***Sparse*** - Similar to other pixels (features) in <u>other images</u> (with respect to smooth transformations)

   Captured by (dense) **image correspondence**

One of these things is not like the others



One of these things is not like the others

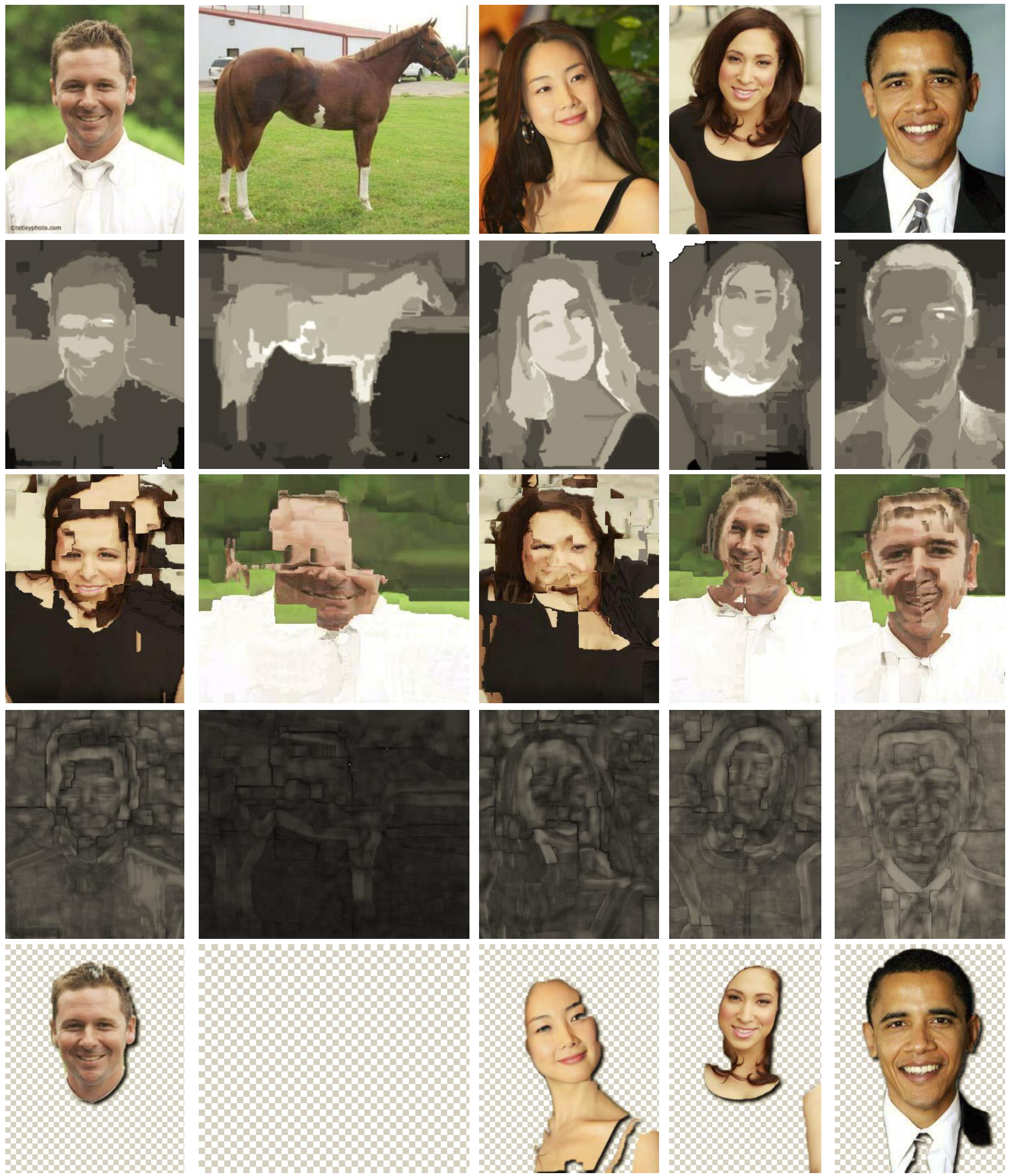
Source

Saliency

Warped neighbor

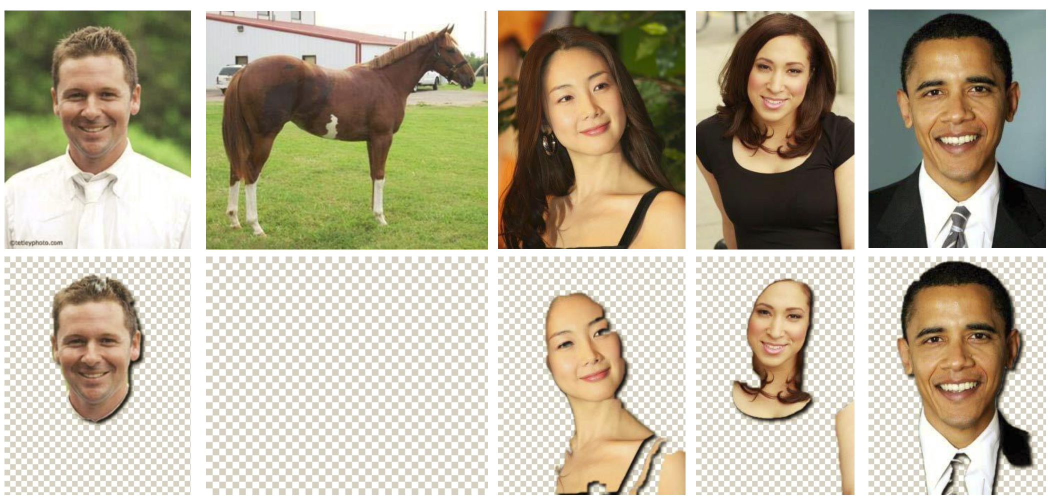
Matching Score

Segmentation



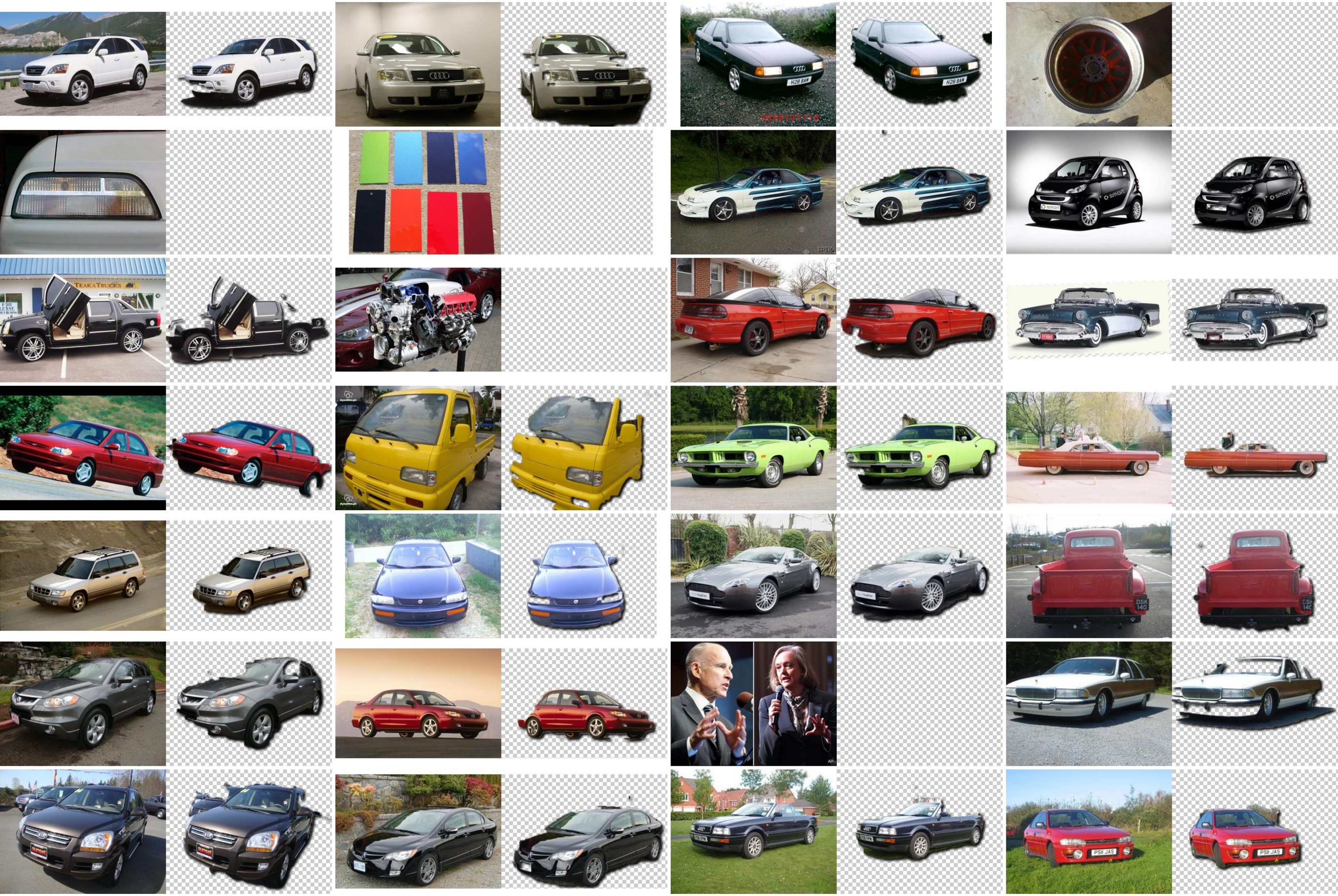
One of these things is not like the others

Horse

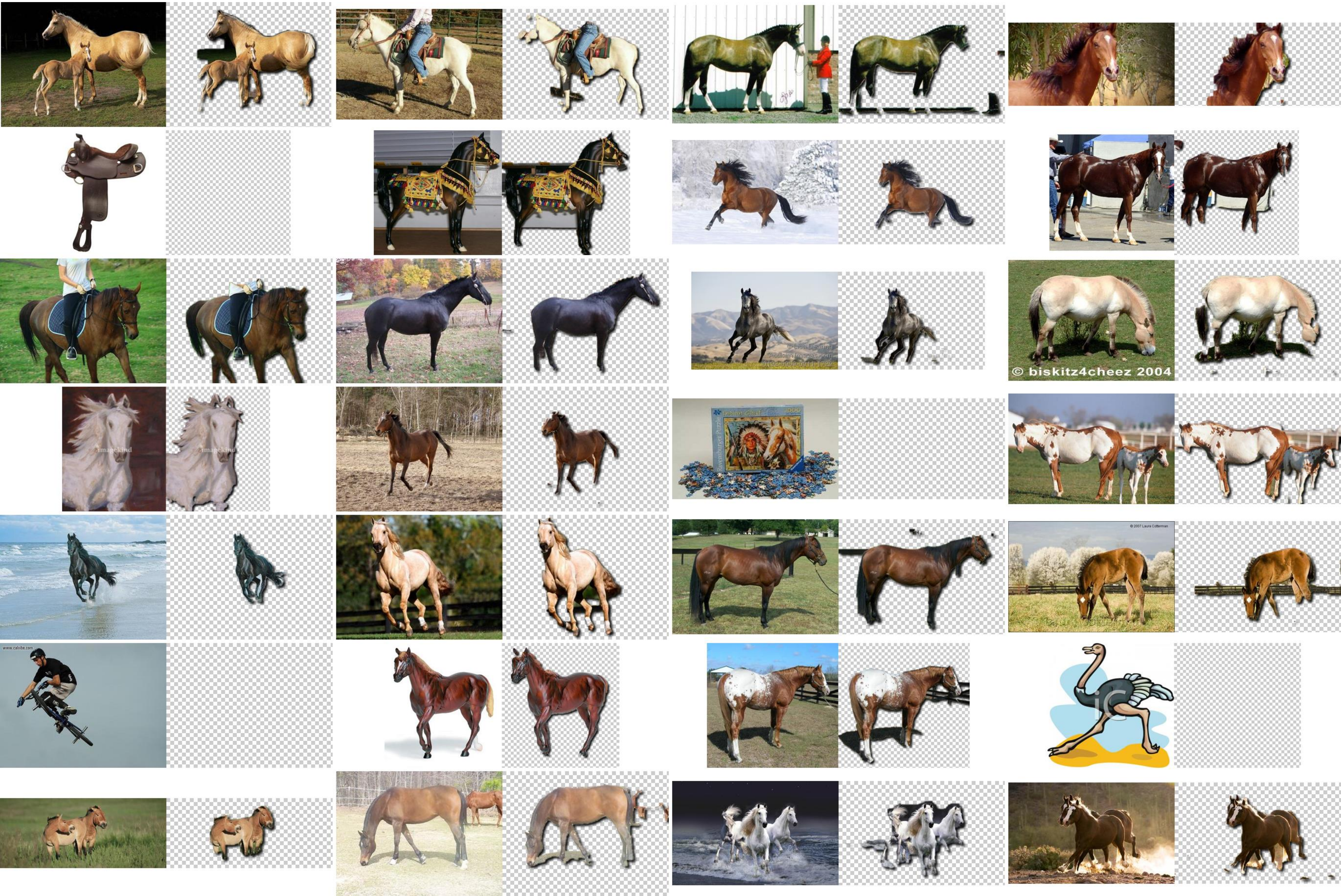


Face

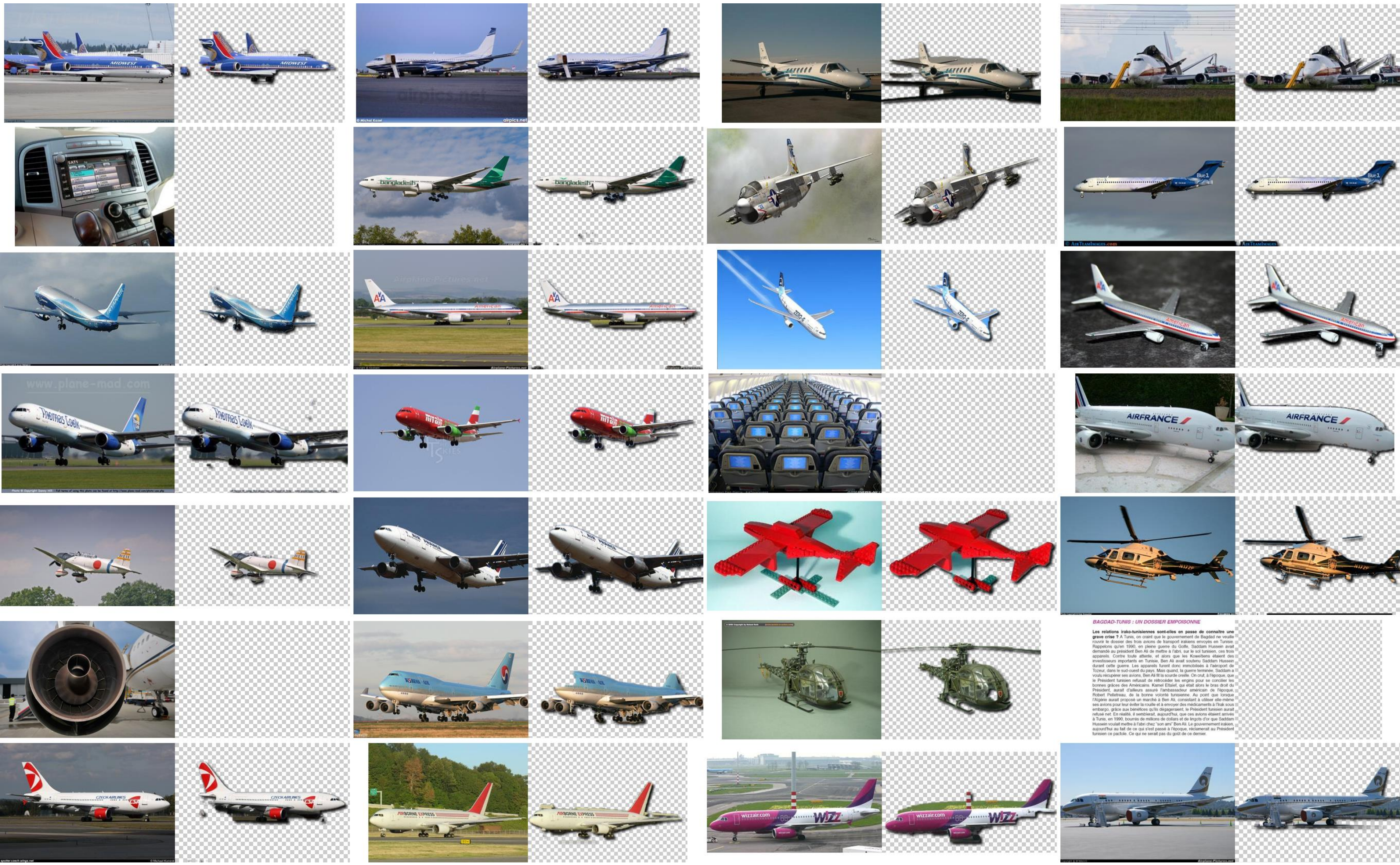
Car (4,347 images, 11% noise)



Horse (6,381 images, 7% noise)

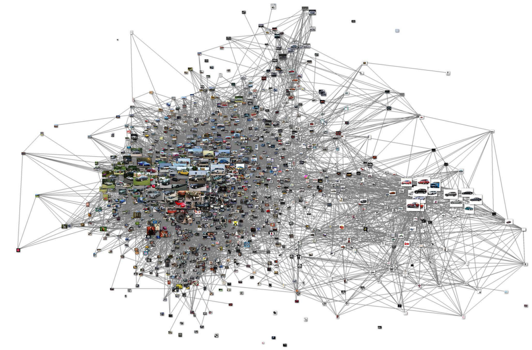


Airplane (4,542 images, 18% noise)



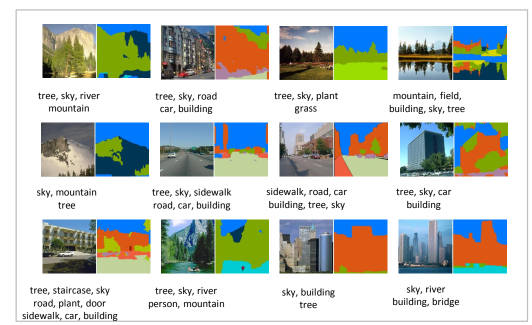
Conclusion

- Labels in big visual data are often unavailable/noisy



- Dense image correspondence (SIFT flow, and others) useful to capture structure, resolve visual ambiguity
  - Becoming a mature technology

- Joint inference for weakly-labeled image databases



  - Annotation Propagation: partial tags + very few (possibly none) pixel labels

  - Object discovery and segmentation: only assuming some underlying "common object"

Thank you!

Michael Rubinstein

MIT CSAIL