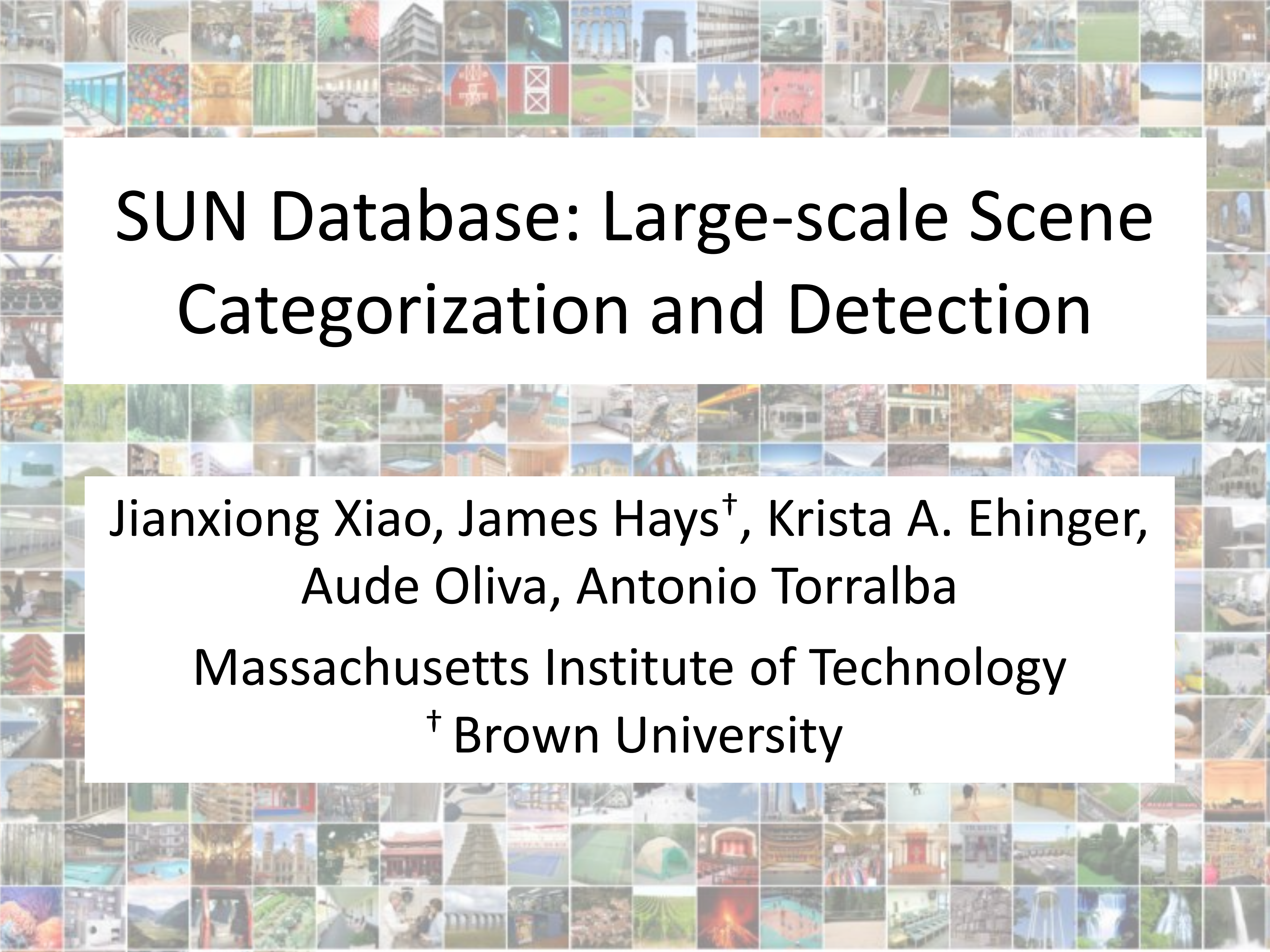
Recap: Viola/Jones detector

- Rectangle features
- Integral images for fast computation
- Boosting for feature selection
- Attentional cascade for fast rejection of negative windows

Project 3

- I have office hours today
- Let’s talk more about scene recognition



SUN Database: Large-scale Scene Categorization and Detection

Jianxiong Xiao, James Hays[†], Krista A. Ehinger, Aude Oliva, Antonio Torralba

Massachusetts Institute of Technology

[†] Brown University

Scene Categorization

Oliva and Torralba, 2001

Coast

Forest

Highway

Inside City

Mountain

Open Country

Street

Tall Building

Fei Fei and Perona, 2005

+

Bedroom

Kitchen

Living Room

Office

Suburb

Lazebnik, Schmid, and Ponce, 2006

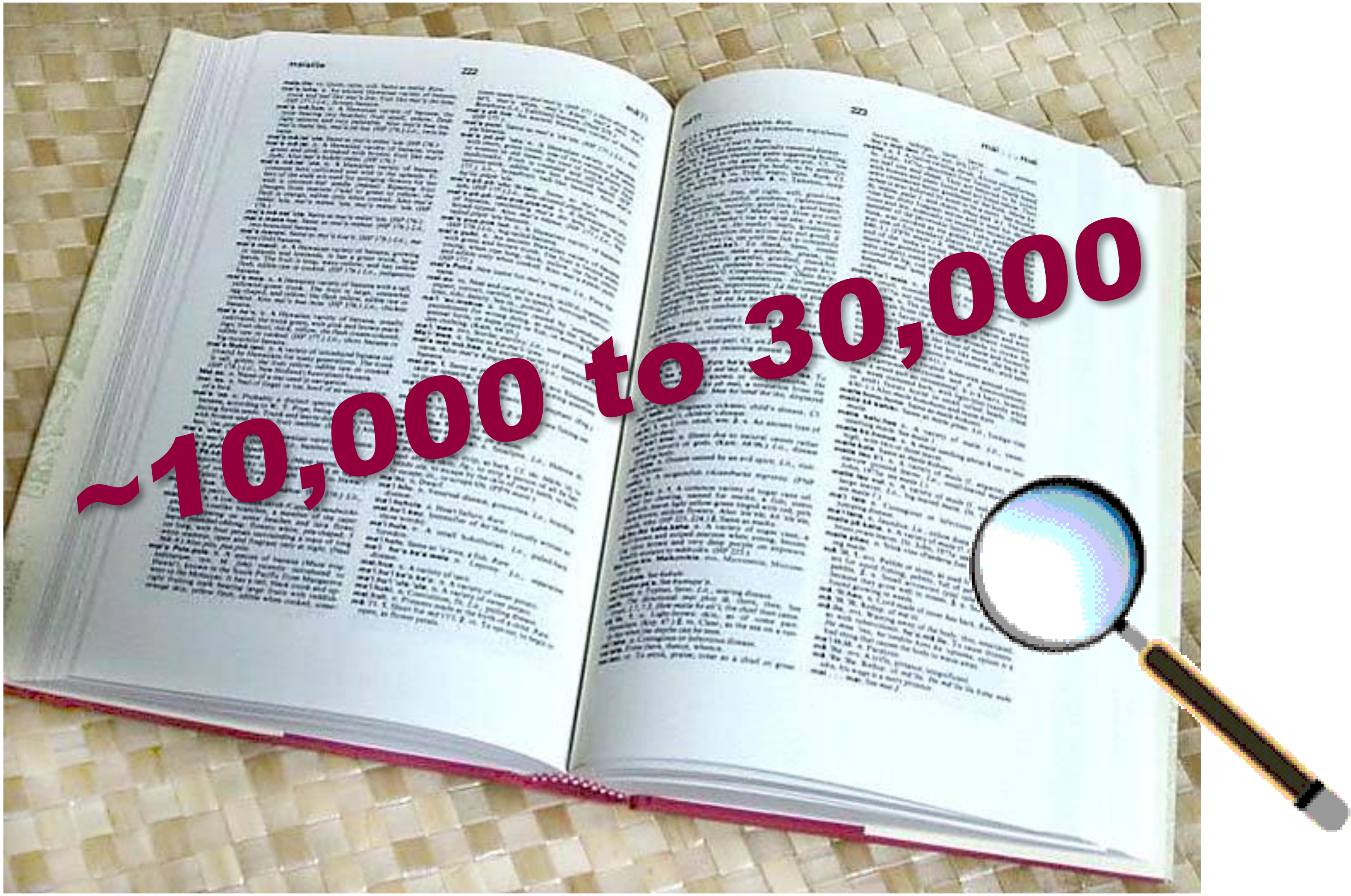
+

Industrial

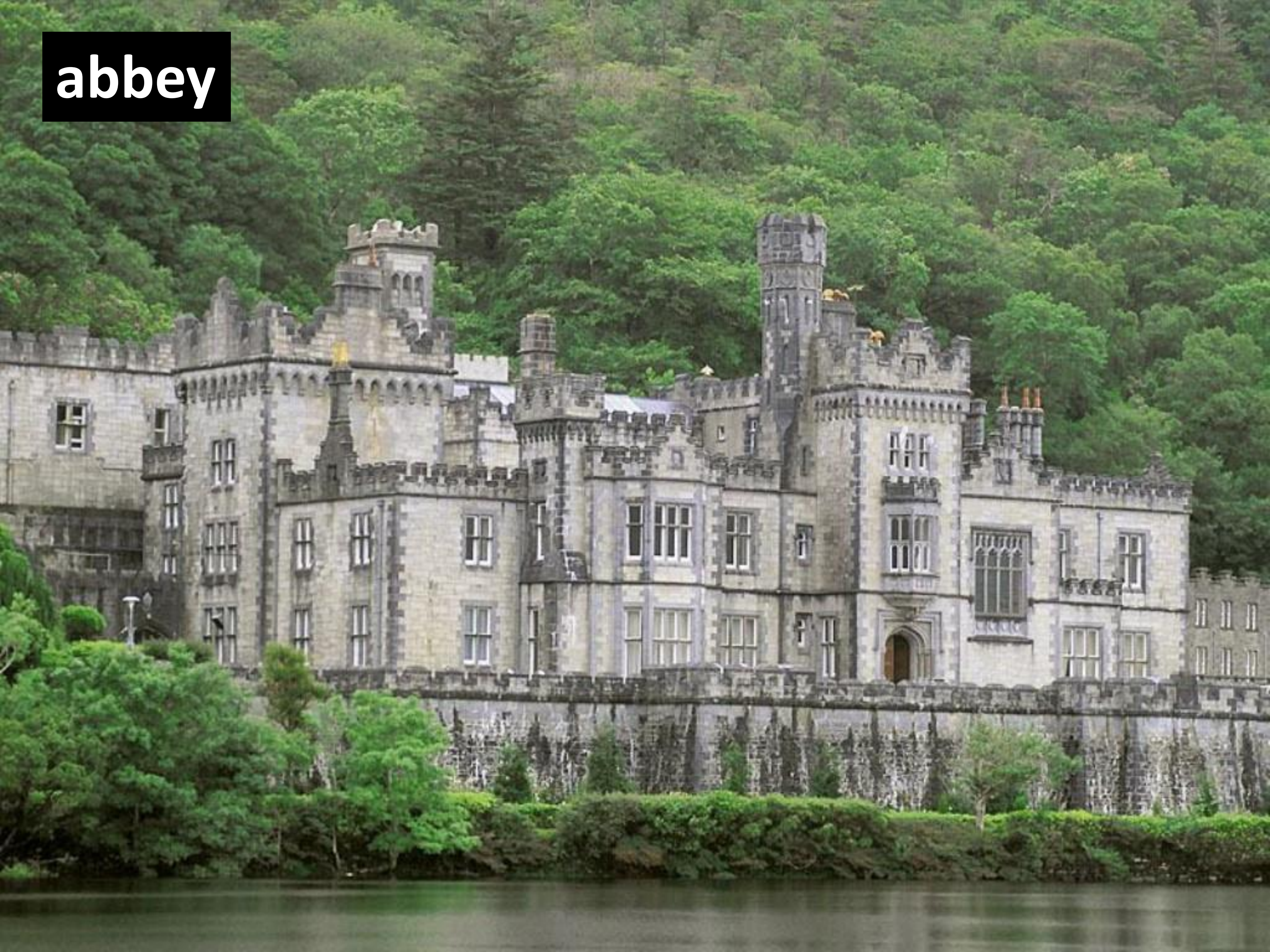
Store

15 Scene Database

How many object categories are there?



Biederman 1987



abbey

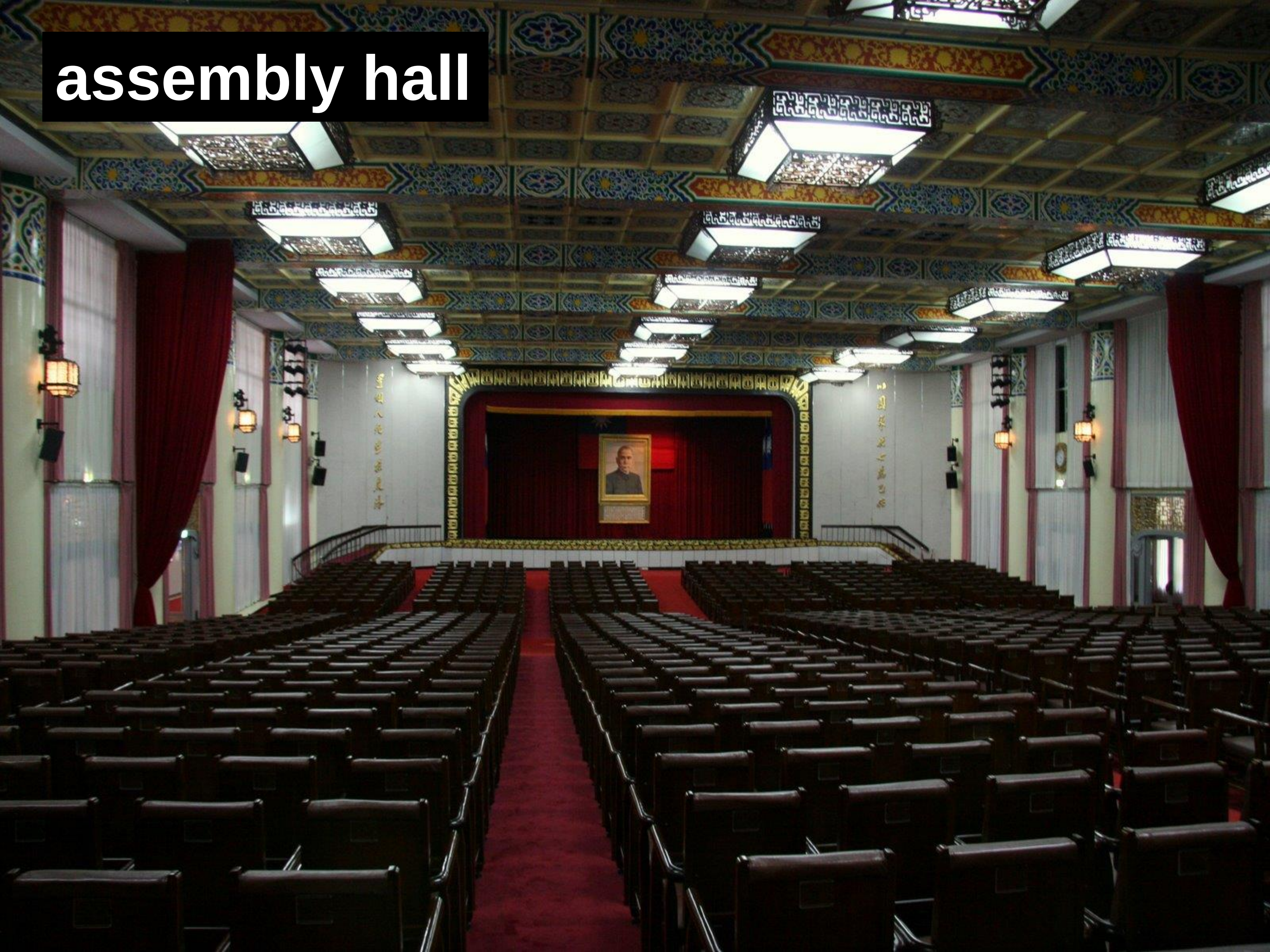
airplane cabin

airport terminal

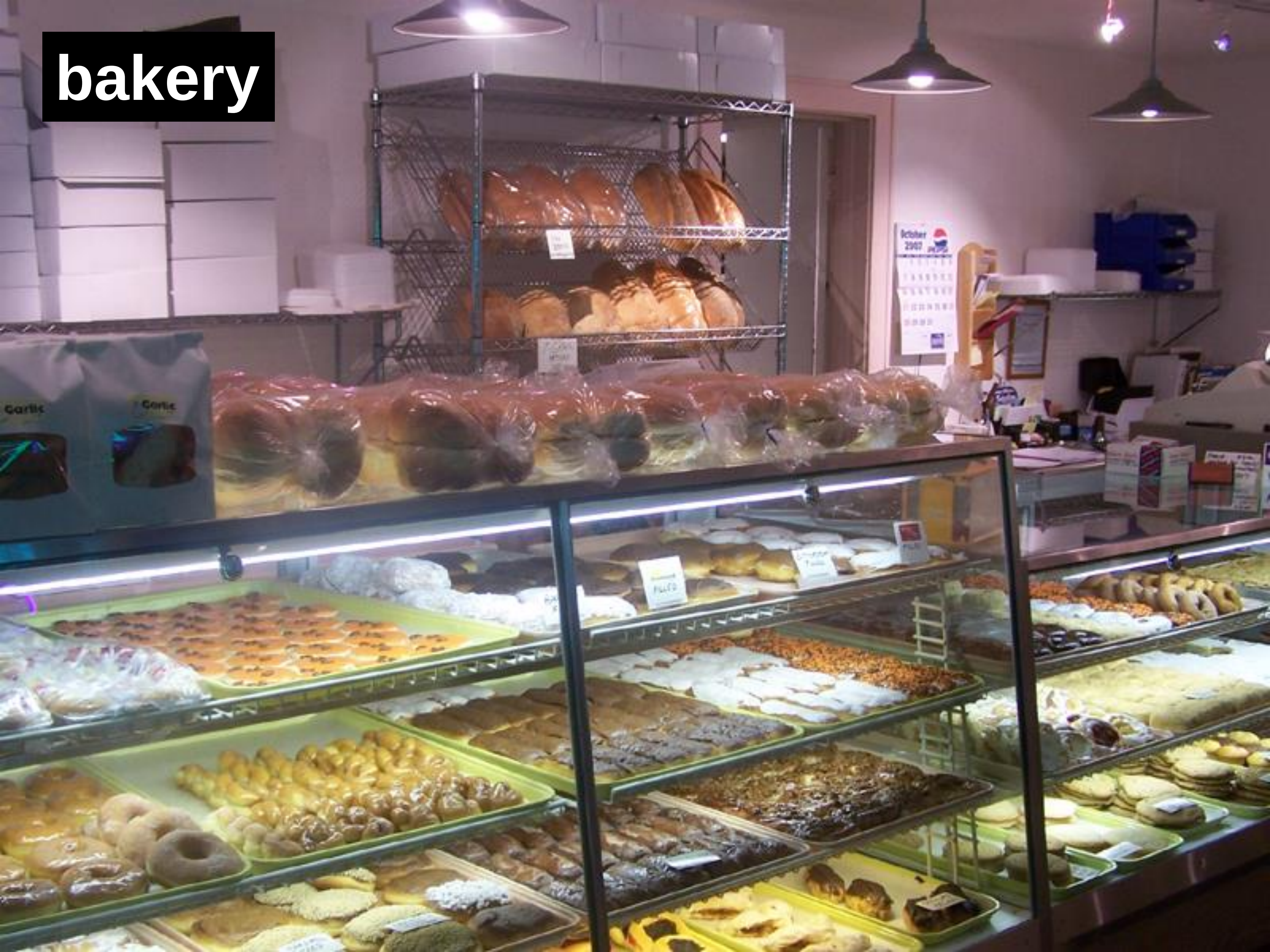
■ ■ ■

apple orchard

assembly hall

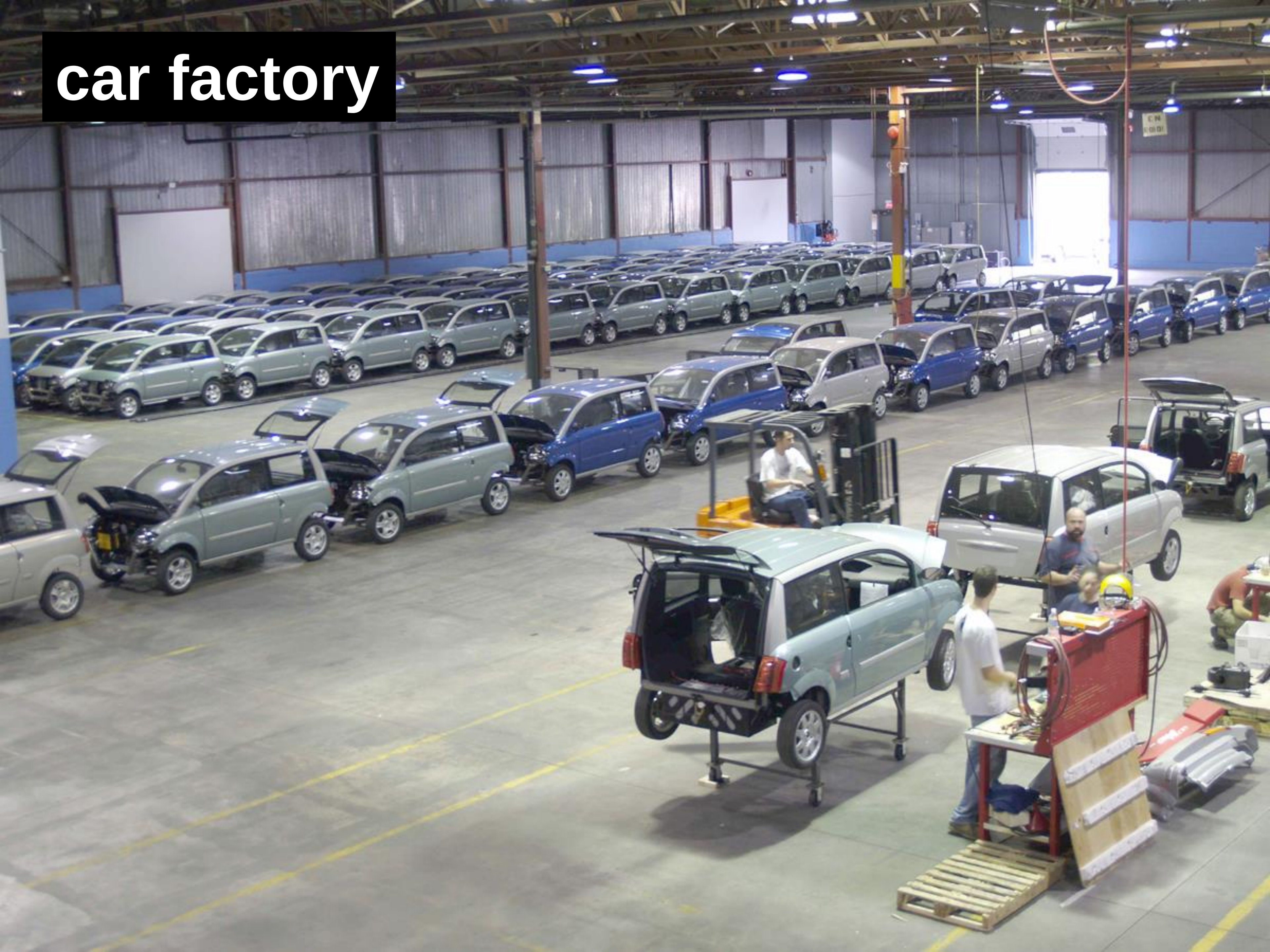


bakery

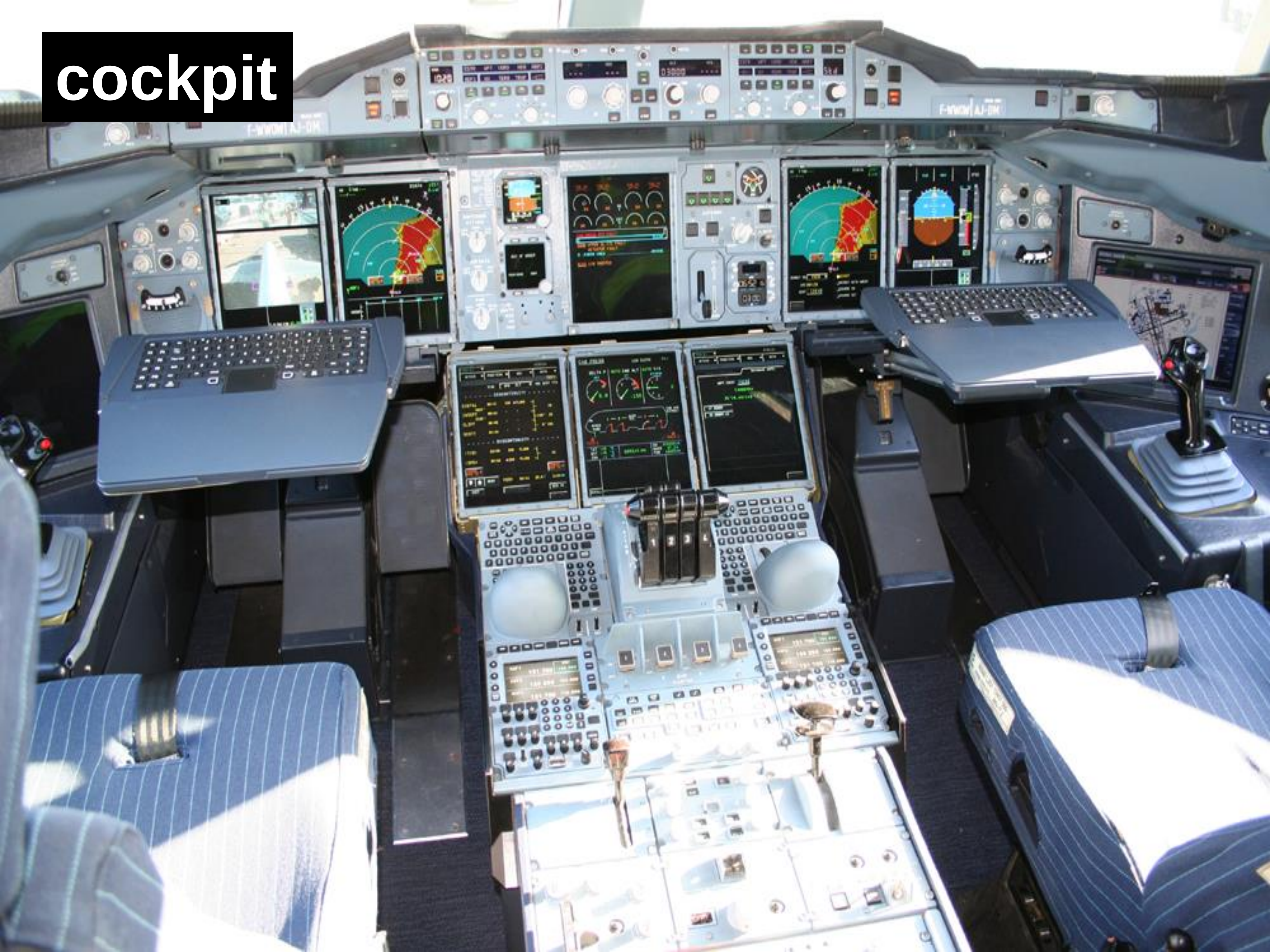


■ ■ ■

car factory



cockpit



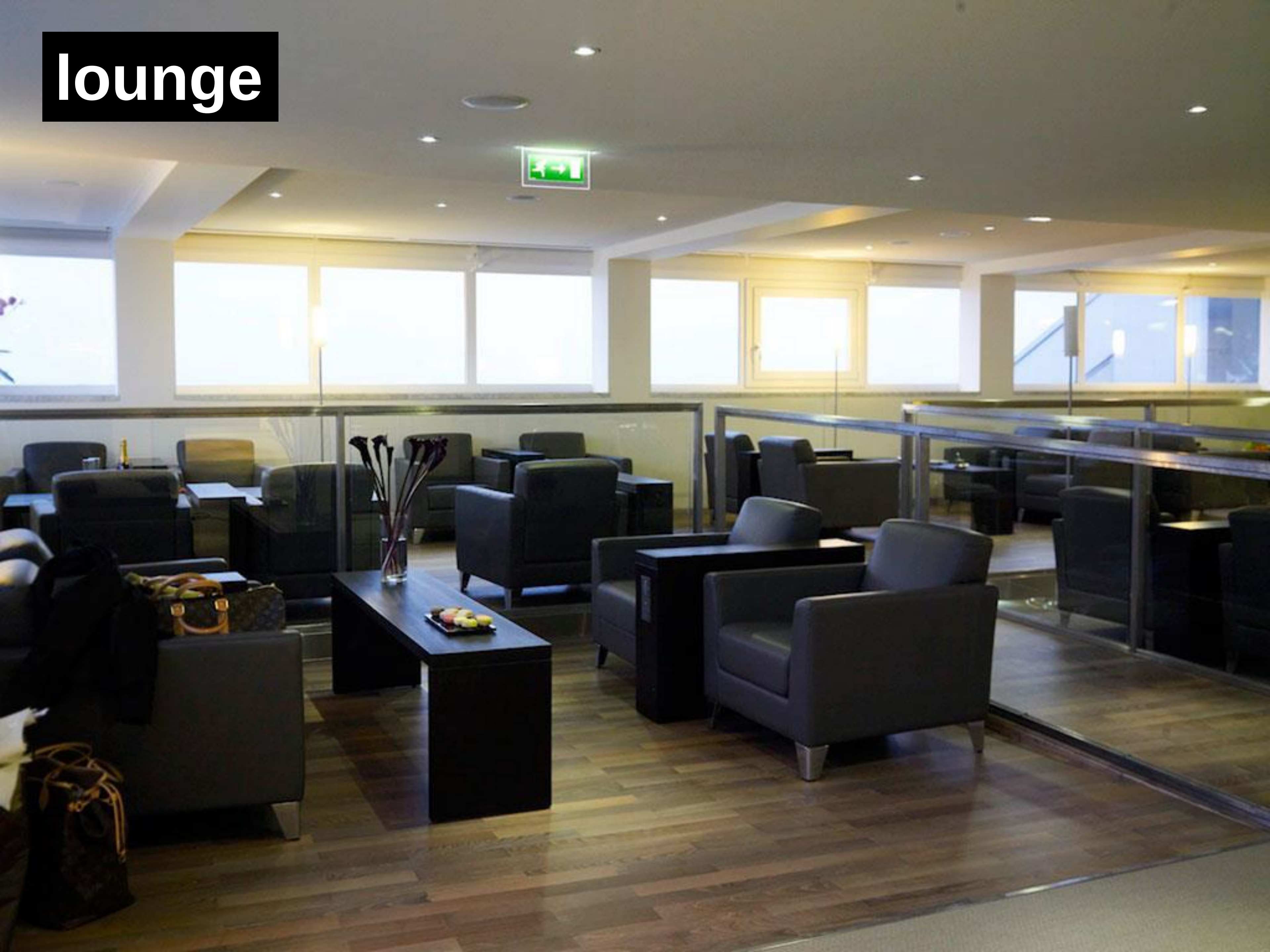
construction site

■ ■ ■

food court

interior car

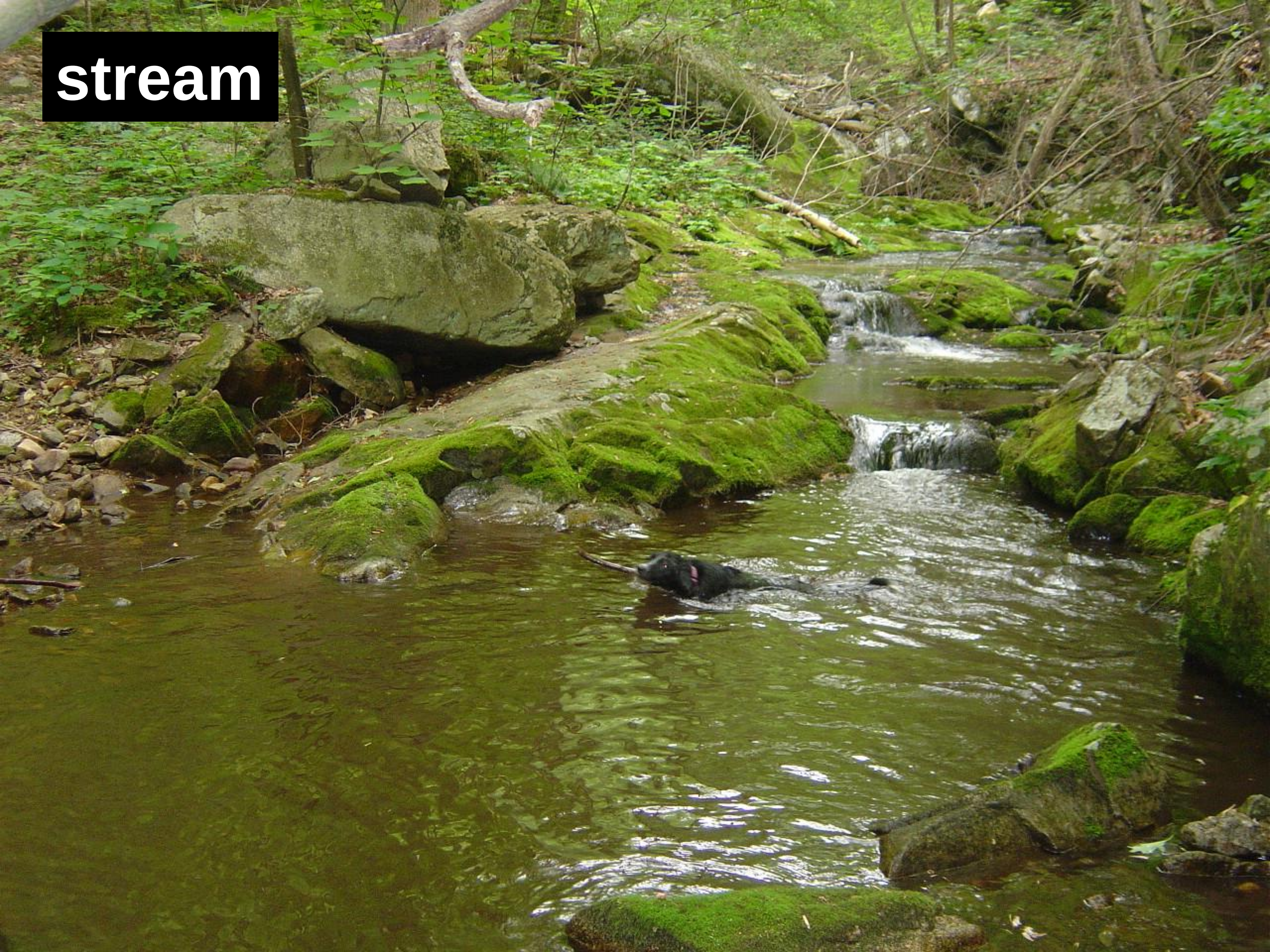
lounge



...

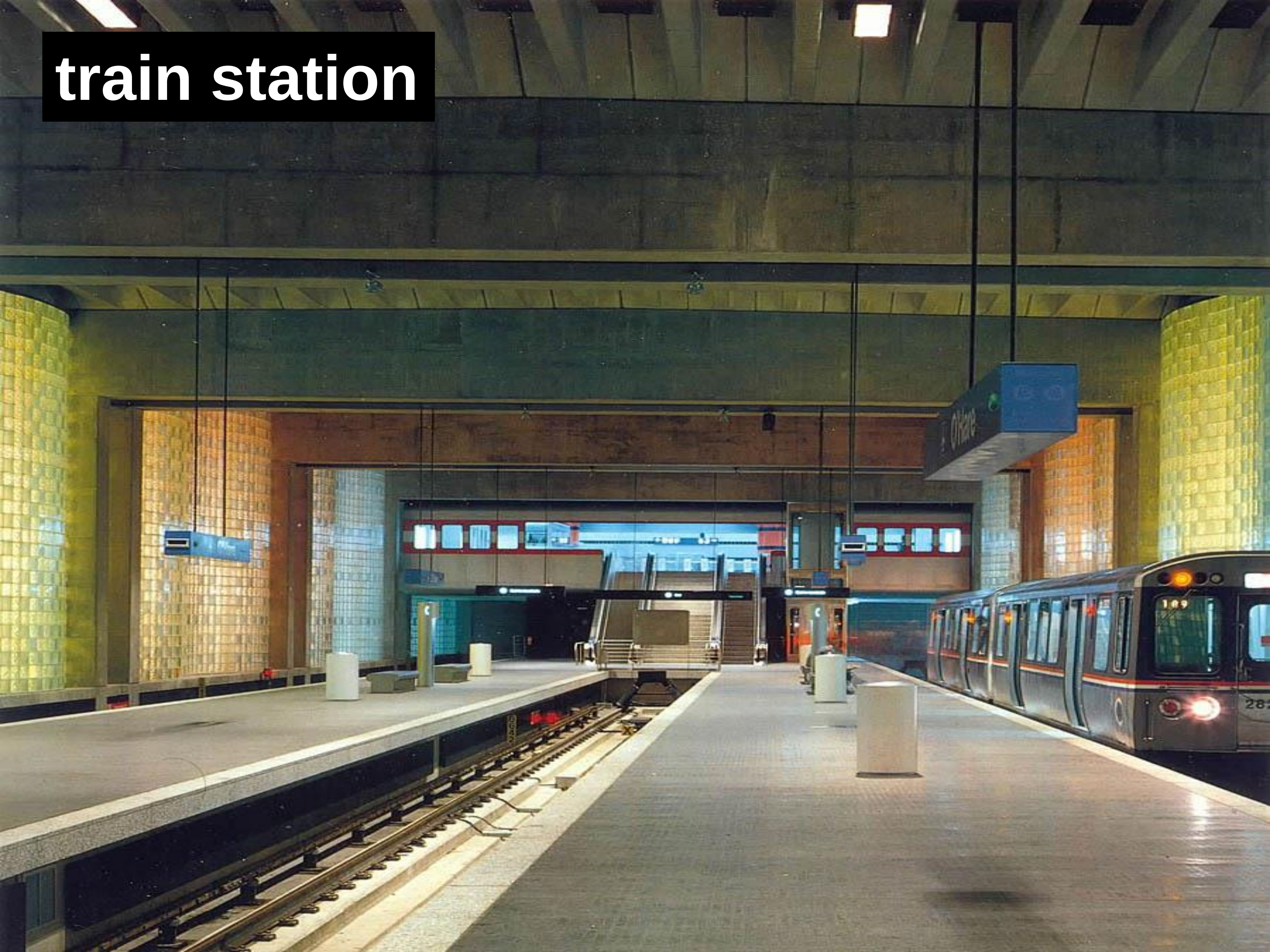
stadium

BLUE SQUARE
BET NOW AT BLUESQUARE.COM
WEMBLEY STADIUM
npower
MAD
VODKAT

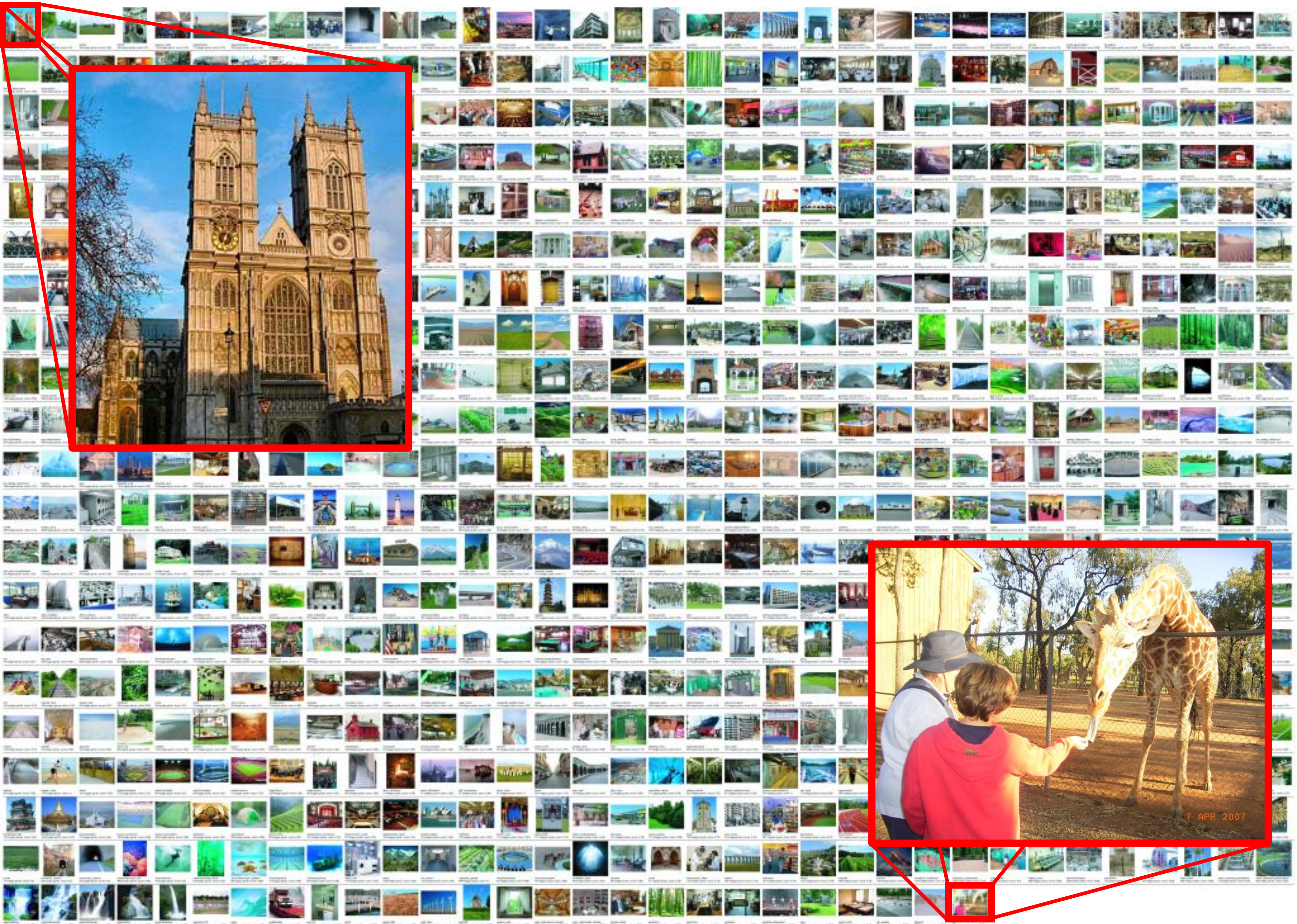


stream

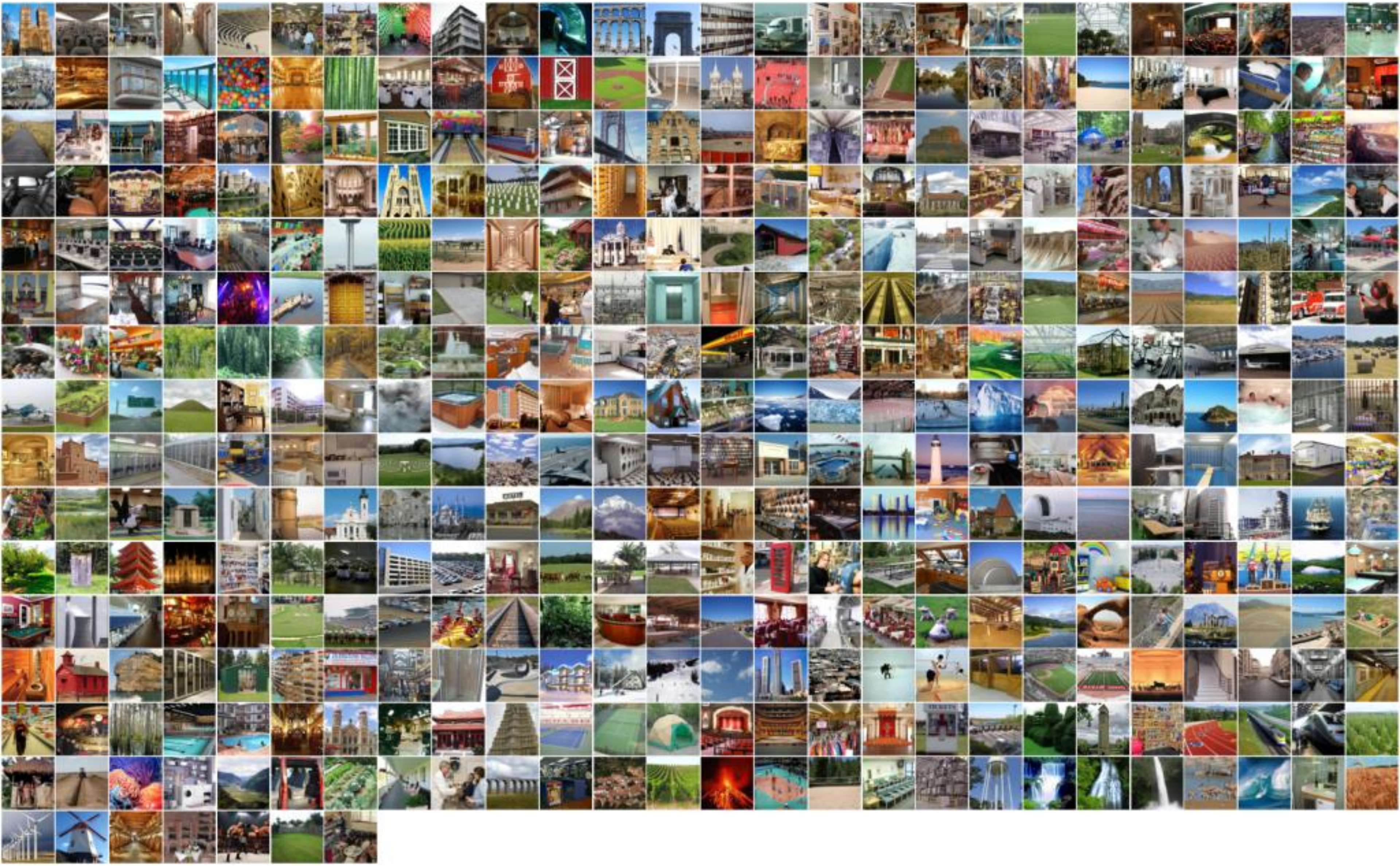
train station



■ ■ ■


7 APR 2007

397 Well-sampled Categories



Evaluating Human Scene Classification

?

"Good worker" Accuracy	98%	90%	68%

bathroom(100%)

beauty salon(100%)

bedroom(100%)

bullring(100%)

playground(100%)

phone booth(100%)

podium outdoor(100%)

greenhouse outdoor(100%)

tennis court outdoor(100%)

wind farm(100%)

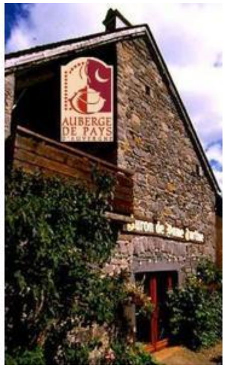
veterinarians office(100%)

riding arena(100%)

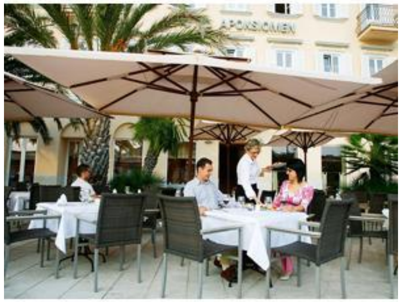
Scene category

Most confusing categories

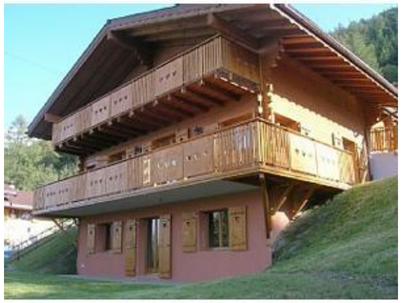
Inn (0%)



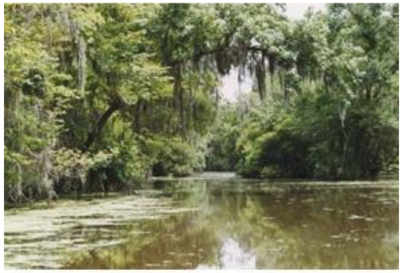
Restaurant patio (44%)



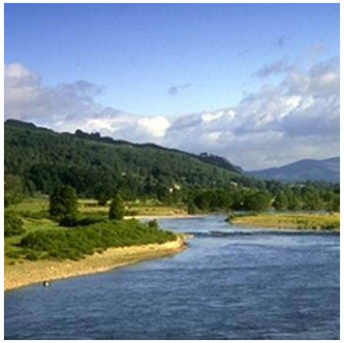
Chalet (19%)



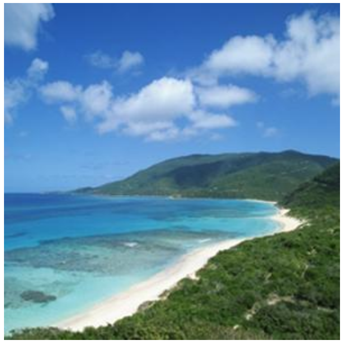
Bayou (0%)



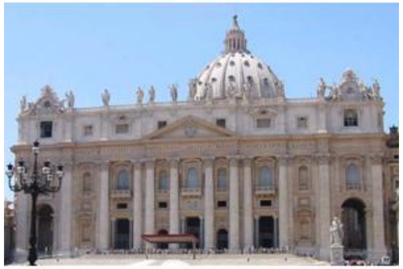
River (67%)



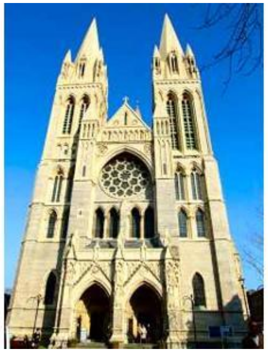
Coast (8%)



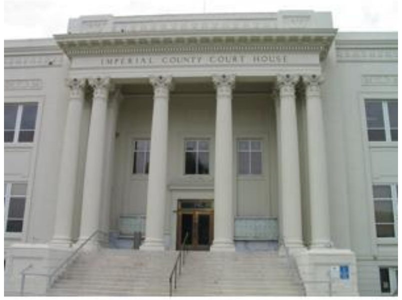
Basilica (0%)



Cathedral(29%)



Courthouse (21%)

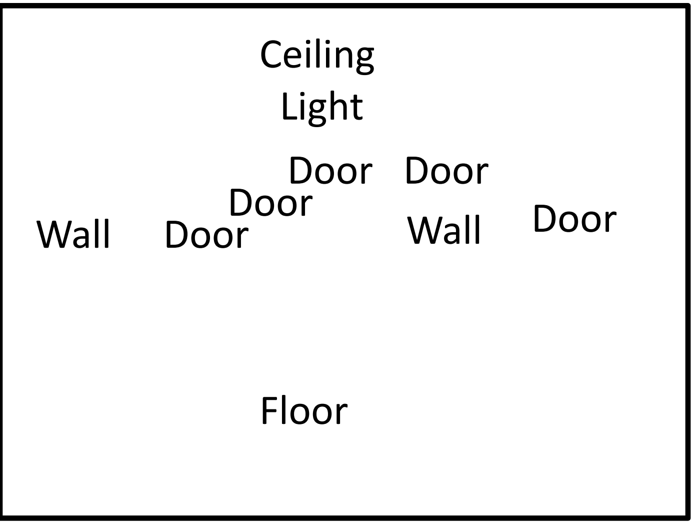


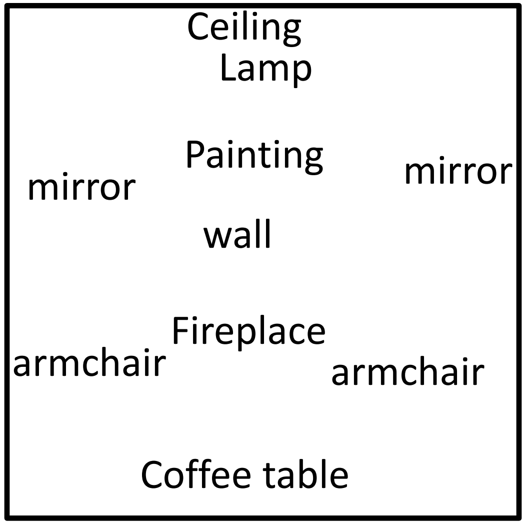
Conclusion: humans can do it

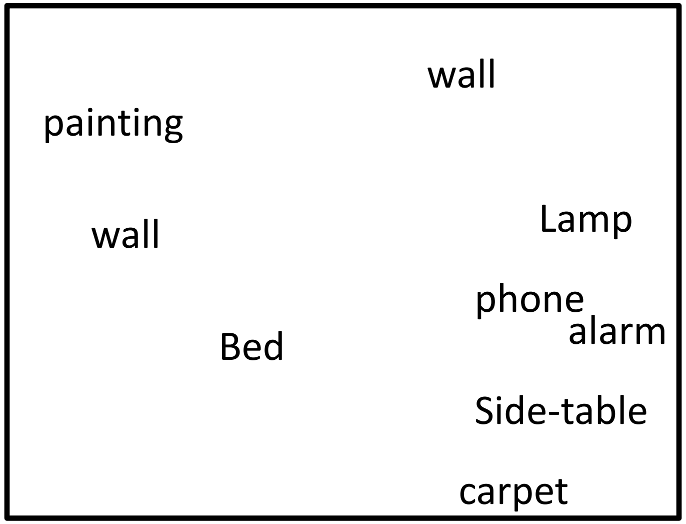
- The SUN database is reasonably consistent and differentiable -- even with a huge number of very specific categories, humans get it right 2/3rds of the time *with no training.*
- We also have a good benchmark for computational methods.

How do we classify scenes?

How do we classify scenes?

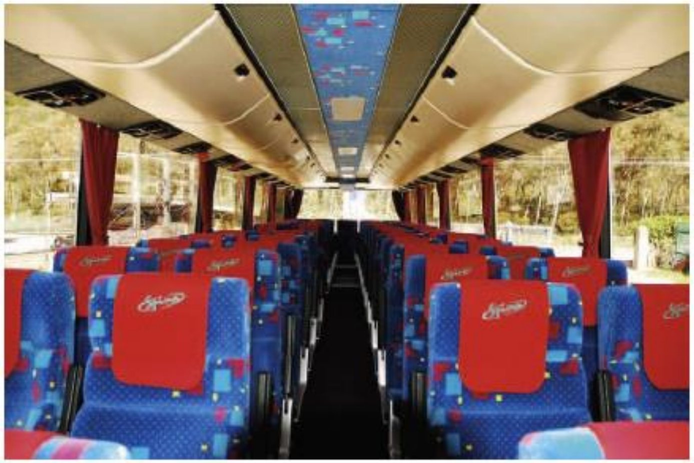


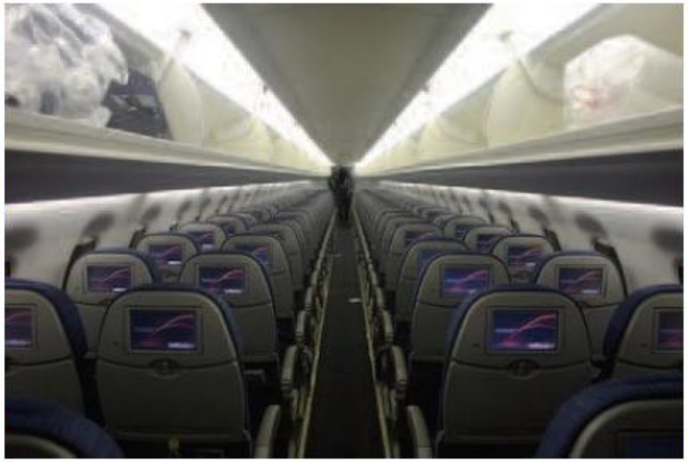


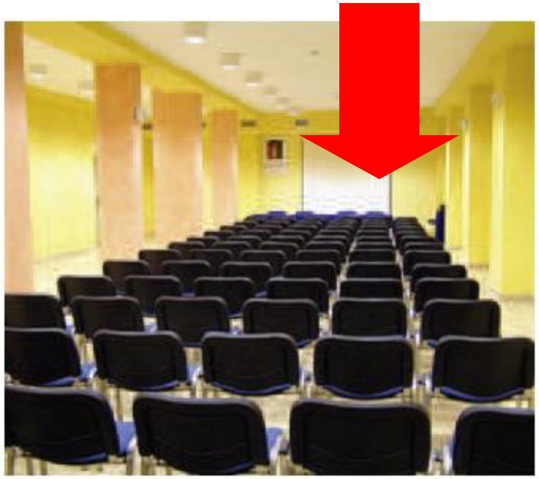


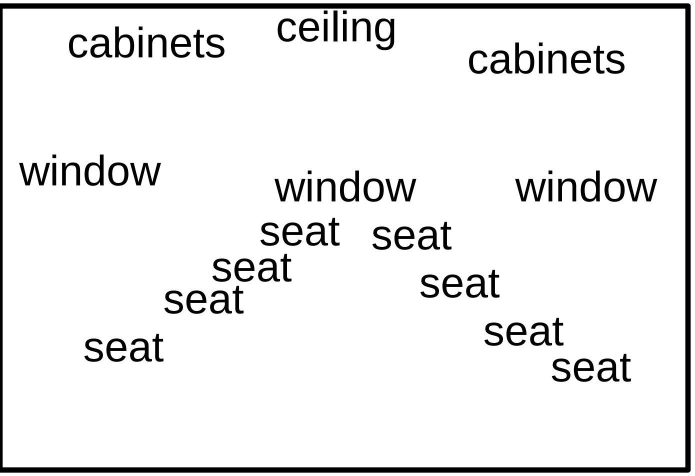
Different objects, different spatial layout

Which are the important elements?









cabinets ceiling cabinets

window window window

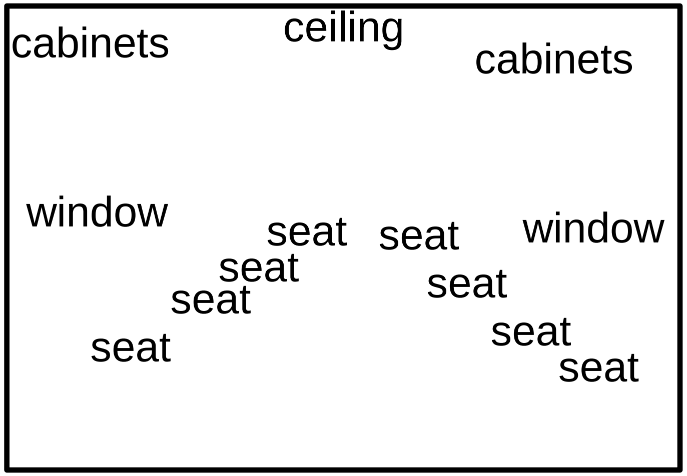
seat seat

seat seat

seat

seat seat

seat



cabinets ceiling cabinets

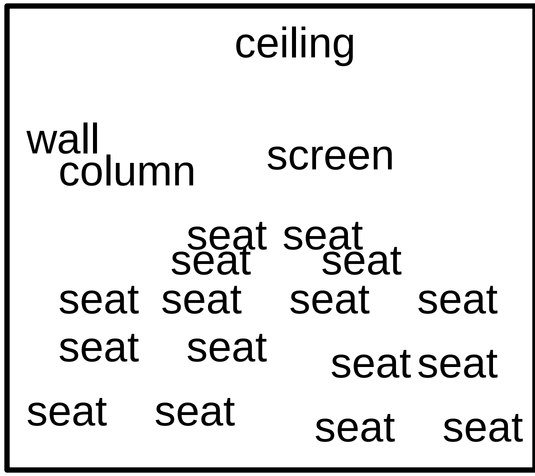
window seat seat window

seat seat

seat

seat seat

seat



ceiling

wall

column screen

seat seat

seat seat

seat seat seat seat

seat seat seat seat

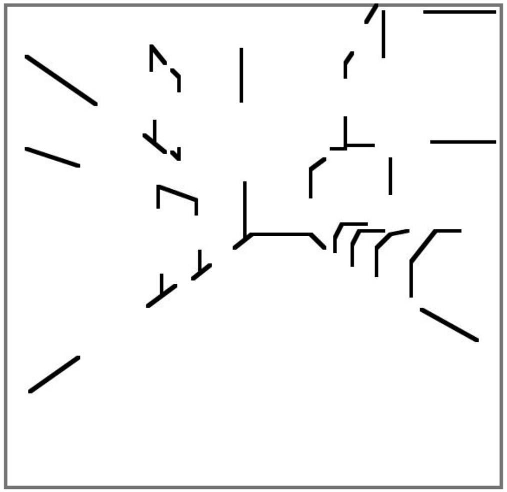
seat seat seat seat

Similar objects, and similar spatial layout

Different lighting, different materials, different “stuff”

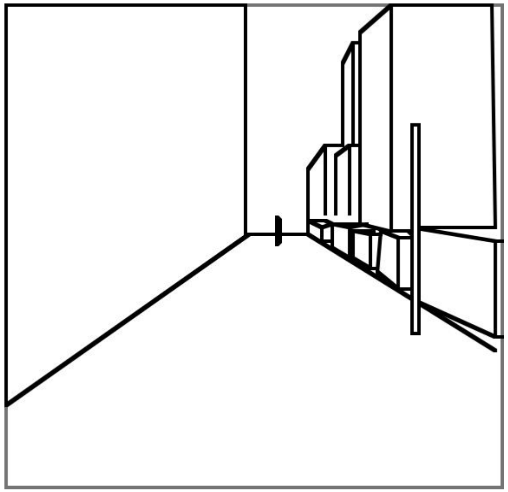
Scene emergent features

“Recognition via features that are not those of individual objects but “emerge” as objects are brought into relation to each other to form a scene.” – Biederman 81



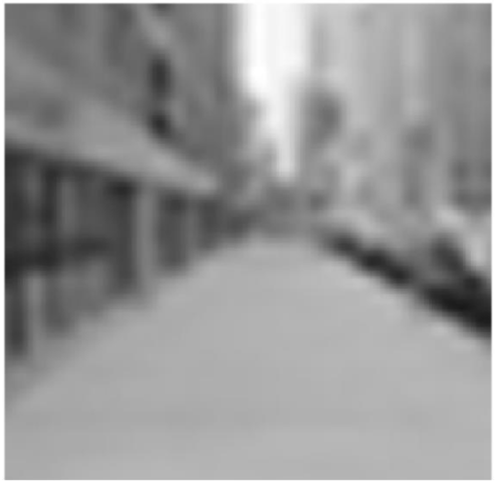
Biederman, 1981

Suggestive edges and junctions



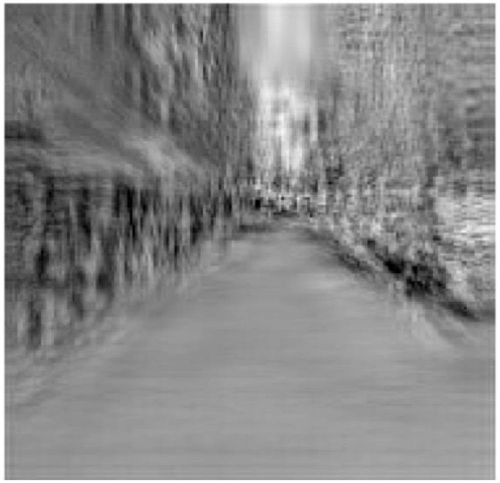
Biederman, 1981

Simple geometric forms



Bruner and Potter, 1969

Blobs



Oliva and Torralba, 2001

Textures

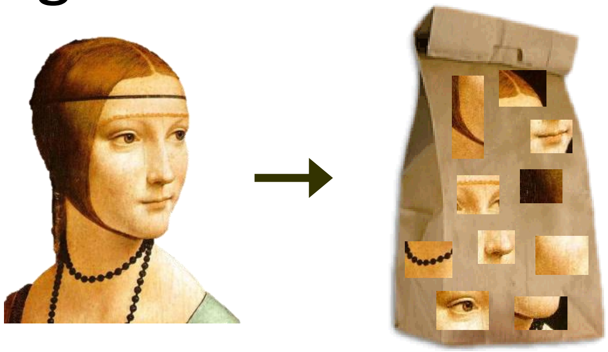
Global Image Descriptors

- Tiny images (Torralba et al, 2008)
- Color histograms
- Self-similarity (Shechtman and Irani, 2007)
- Geometric class layout (Hoiem et al, 2005)
- Geometry-specific histograms (Lalonde et al, 2007)
- Dense and Sparse SIFT histograms
- Berkeley texton histograms (Martin et al, 2001)
- HoG 2x2 spatial pyramids
- Gist scene descriptor (Oliva and Torralba, 2008)

Texture Features

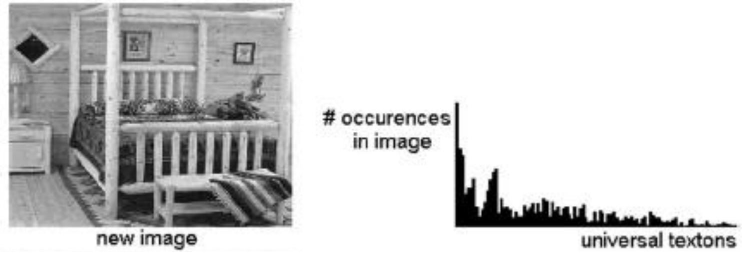
Global Texture Descriptors

Bag of words



Sivic et. al., ICCV 2005
Fei-Fei and Perona, CVPR 2005

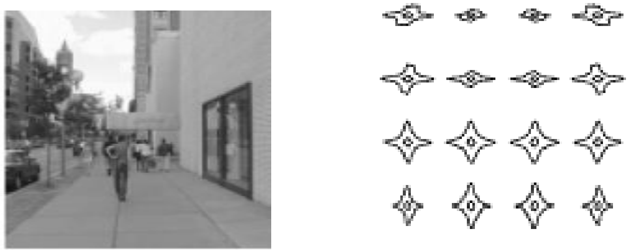
Non localized textons



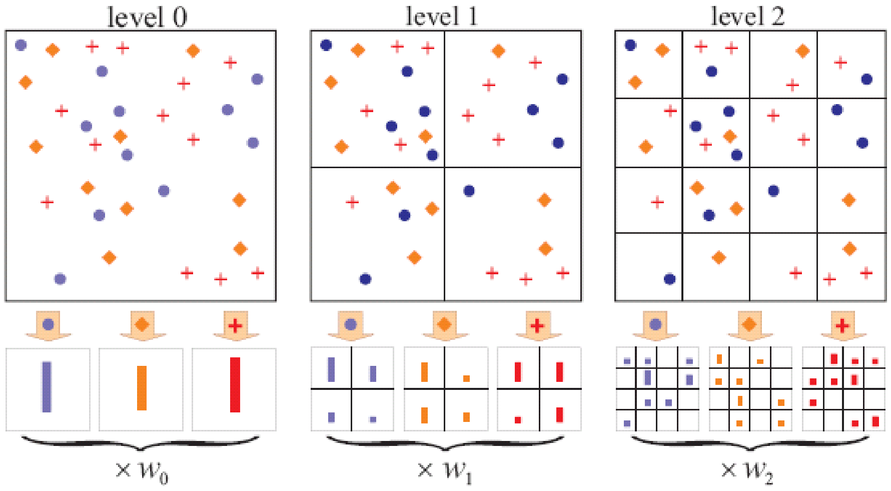
Walker, Malik. Vision Research 2004

…

Spatially organized textures



M. Gorkani, R. Picard, ICPR 1994
A. Oliva, A. Torralba, IJCV 2001



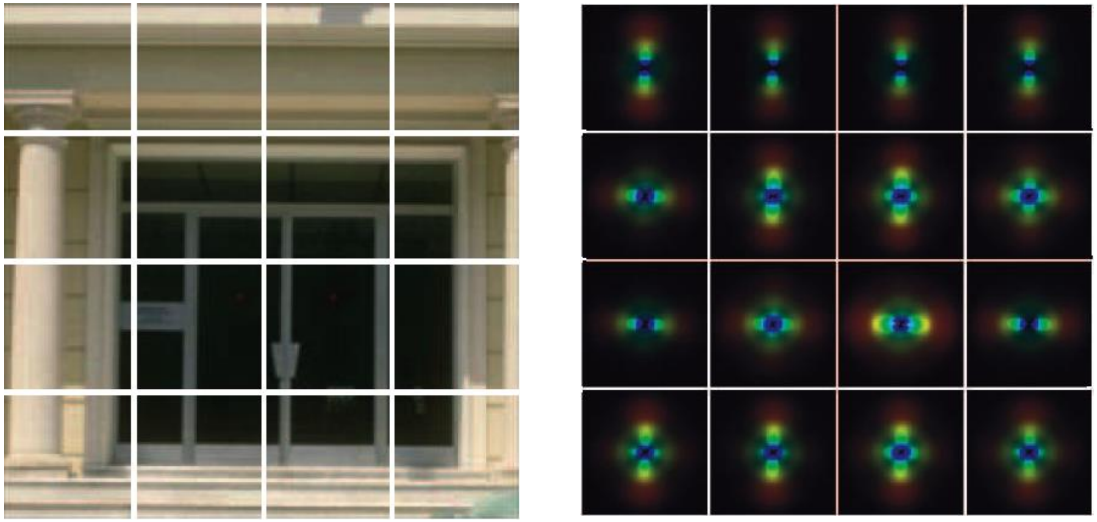
S. Lazebnik, et al, CVPR 2006

…

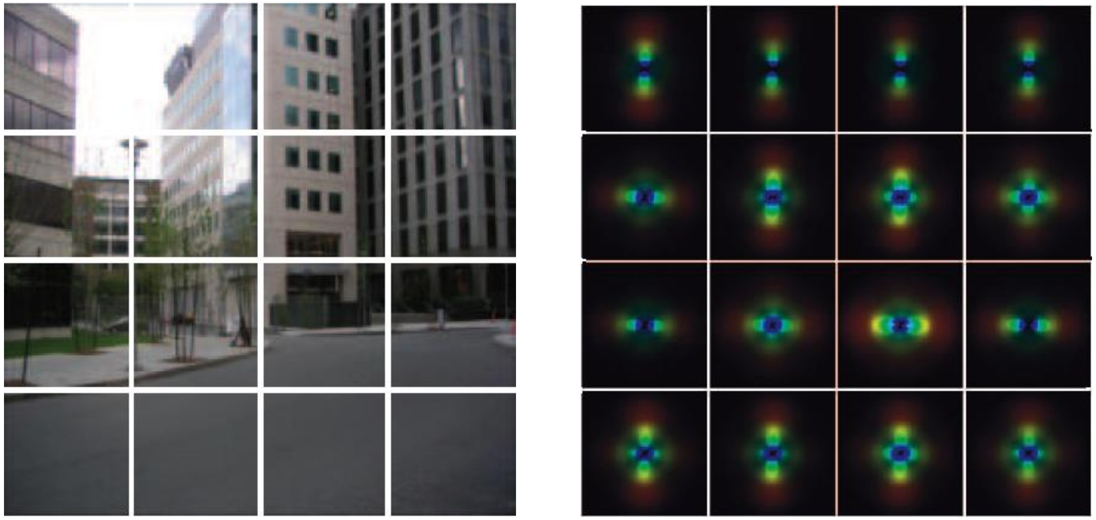
R. Datta, D. Joshi, J. Li, and J. Z. Wang, **Image Retrieval: Ideas, Influences, and Trends of the New Age**, *ACM Computing Surveys*, vol. 40, no. 2, pp. 5:1-60, 2008.

Gist descriptor

Oliva and Torralba, 2001



- Apply oriented Gabor filters over different scales
- Average filter energy in each bin

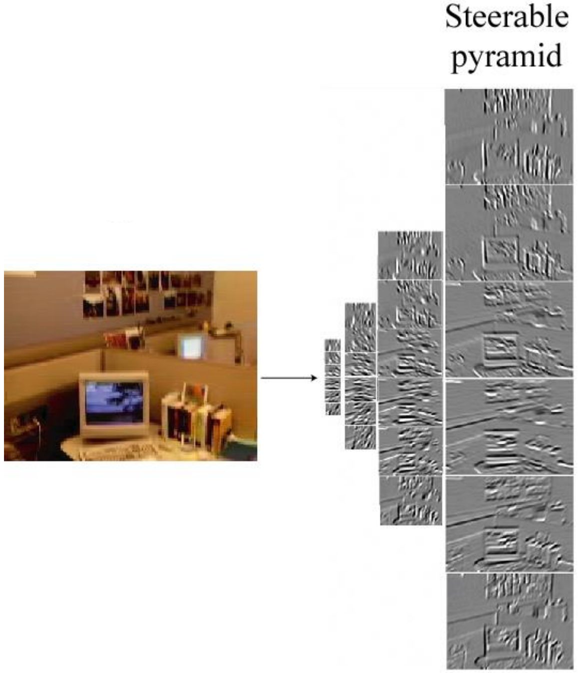


8 orientations
4 scales
x 16 bins
512 dimensions

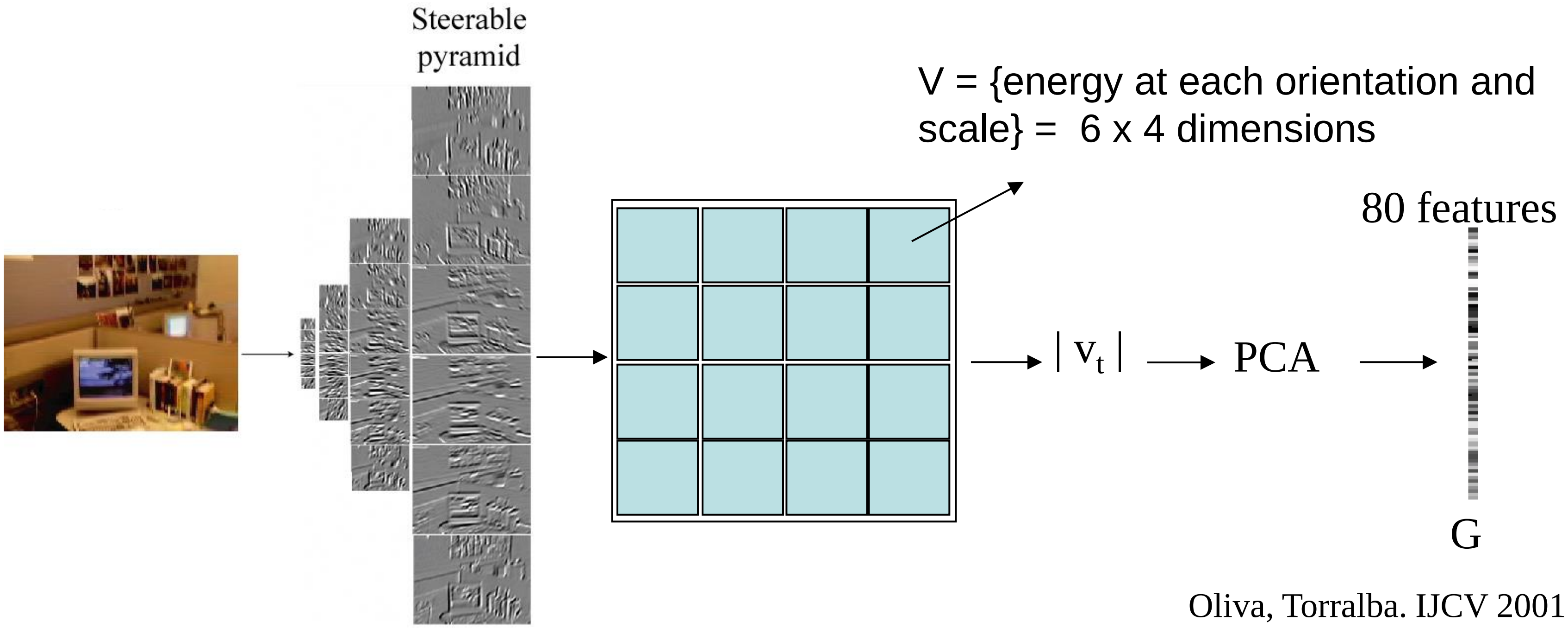
Similar to SIFT (Lowe 1999) applied to the entire image

M. Gorkani, R. Picard, ICPR 1994; Walker, Malik. Vision Research 2004; Vogel et al. 2004;
Fei-Fei and Perona, CVPR 2005; S. Lazebnik, et al, CVPR 2006; …

Gist descriptor



Gist descriptor



Steerable pyramid

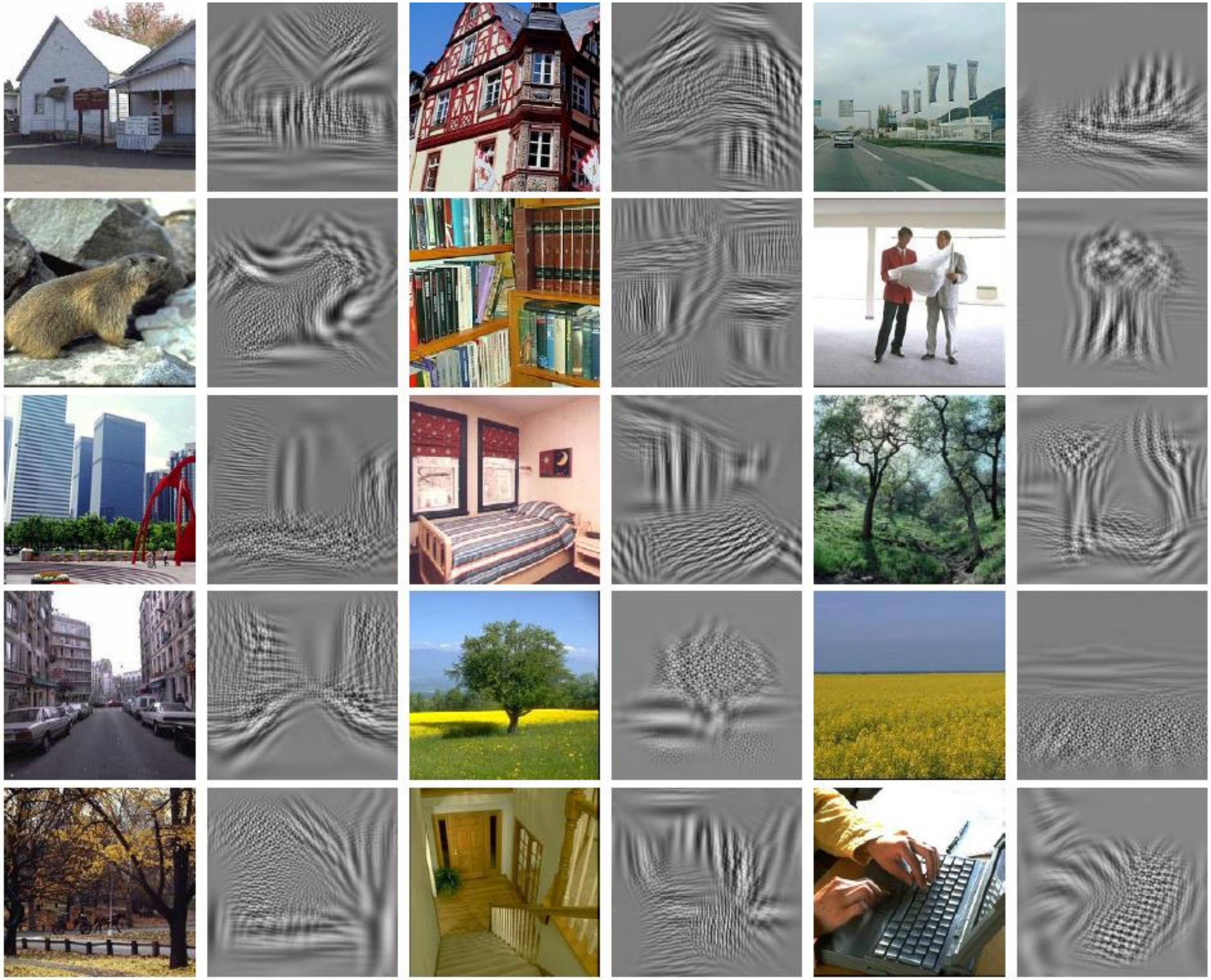
V = {energy at each orientation and scale} = 6 x 4 dimensions

80 features

$|v_t|$ → PCA → G

Oliva, Torralba. IJCV 2001

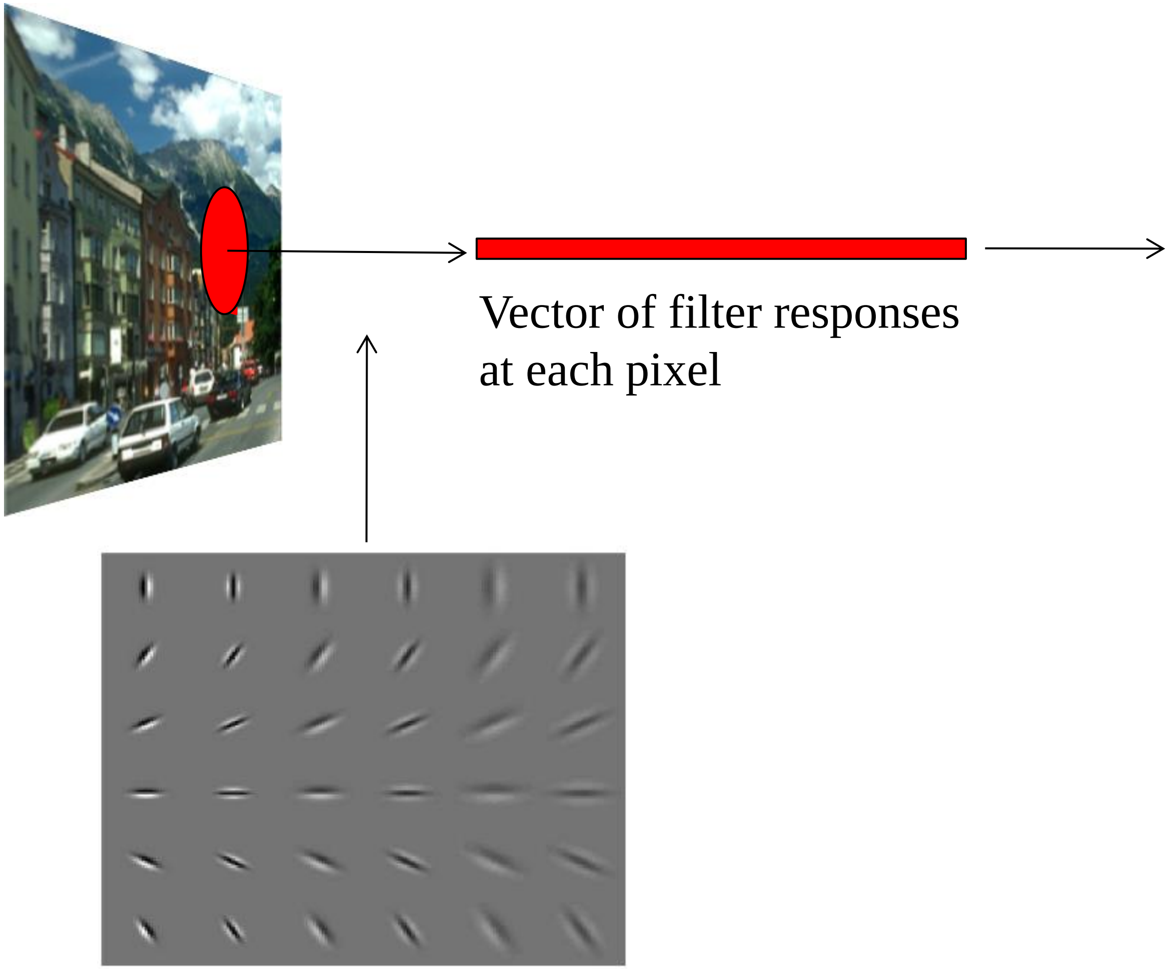
Example visual gists



Global features (I) ~ global features (I')

Oliva & Torralba (2001)

Textons



Vector of filter responses at each pixel

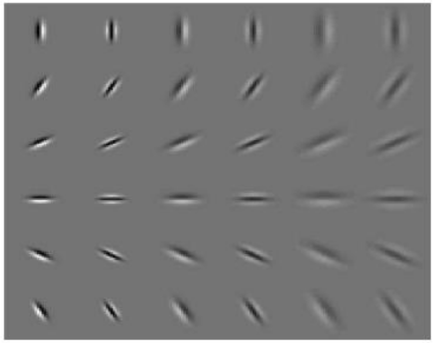
Kmeans over a set of vectors on a collection of images

Filter bank

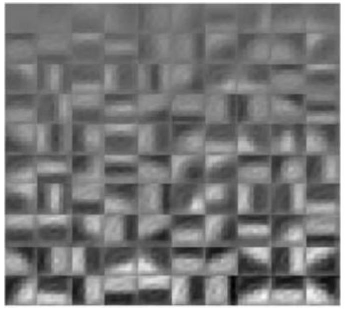
Malik, Belongie, Shi, Leung, 1999

Textons

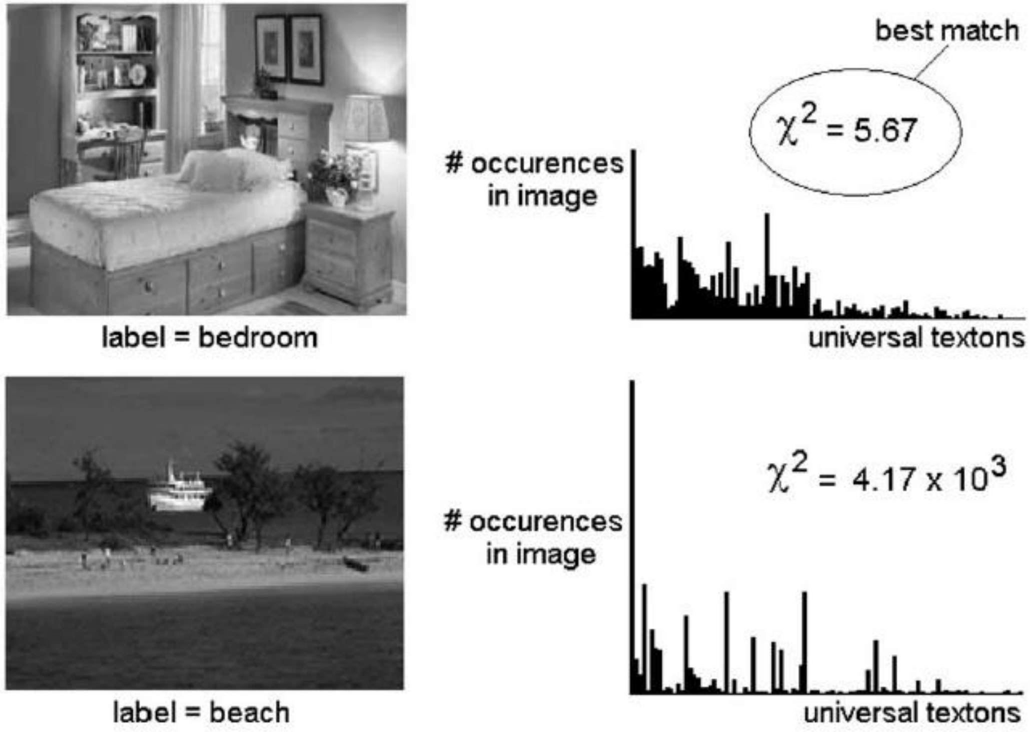
Filter bank



K-means (100 clusters)



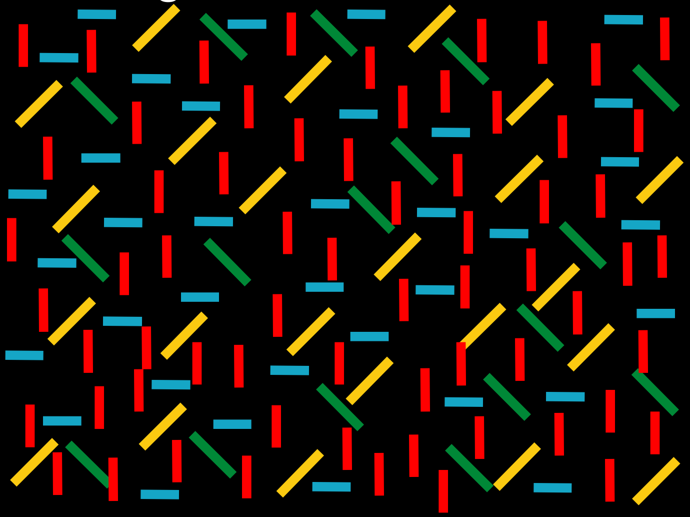
Malik, Belongie, Shi, Leung, 1999

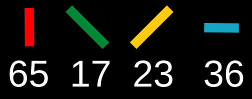


Walker, Malik, 2004

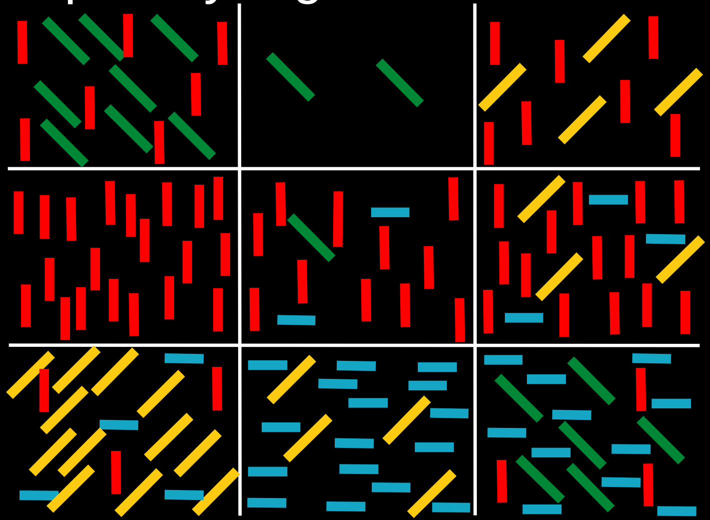
Bag of words

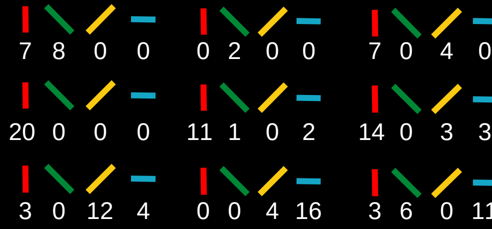
Bag of words model





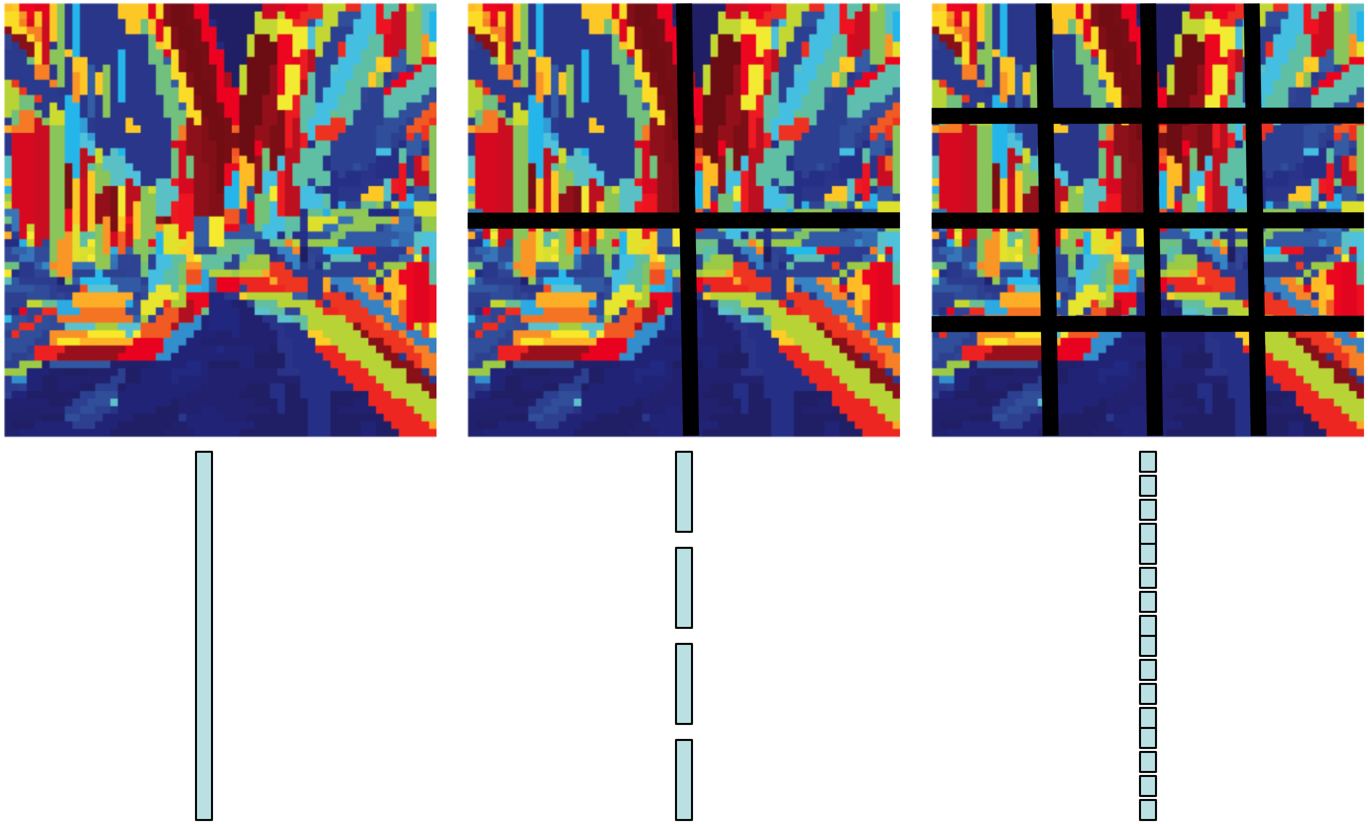
Spatially organized textures





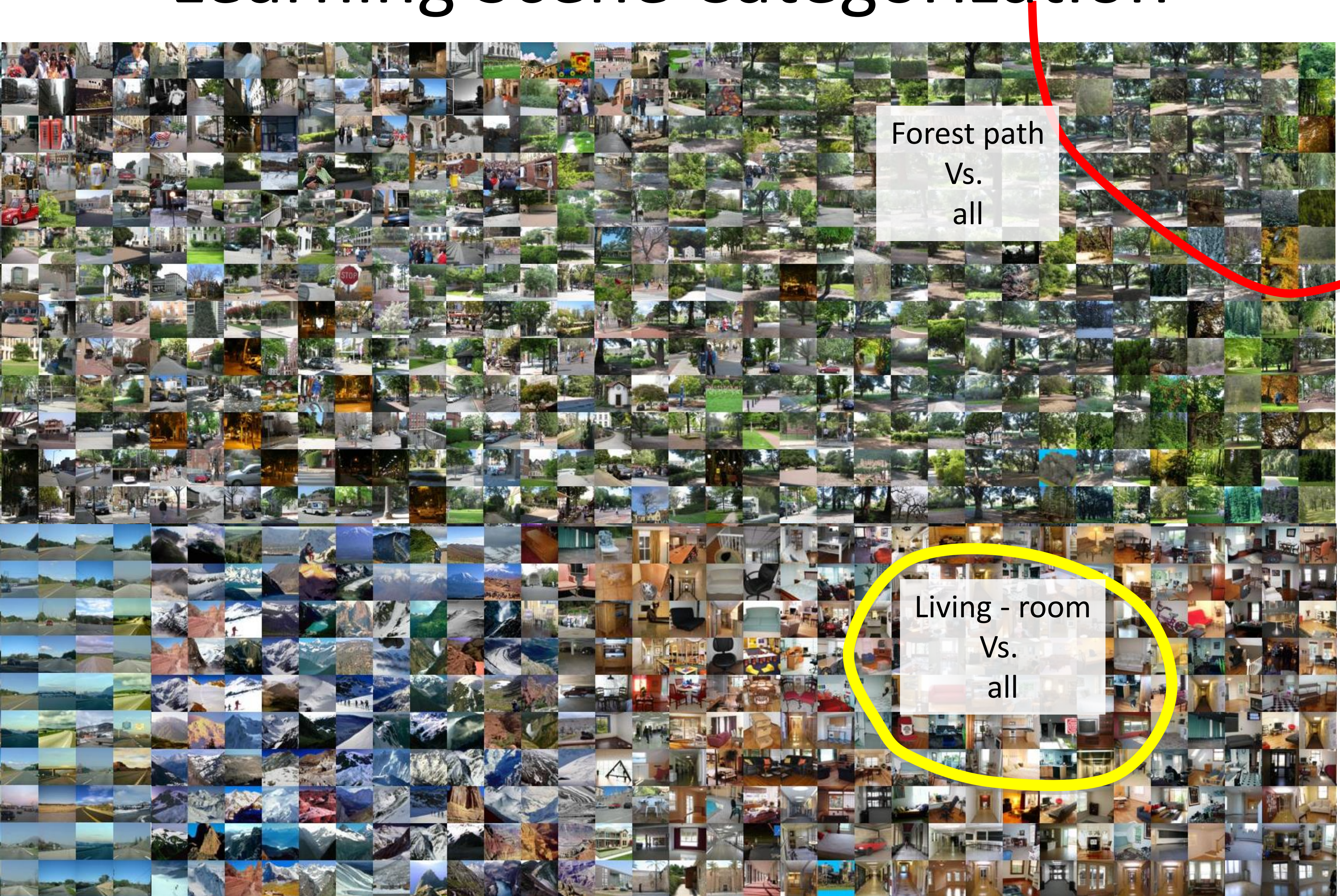
Bag of words & spatial pyramid matching

Sivic, Zisserman, 2003. Visual words = Kmeans of SIFT descriptors



S. Lazebnik, et al, CVPR 2006

Learning Scene Categorization



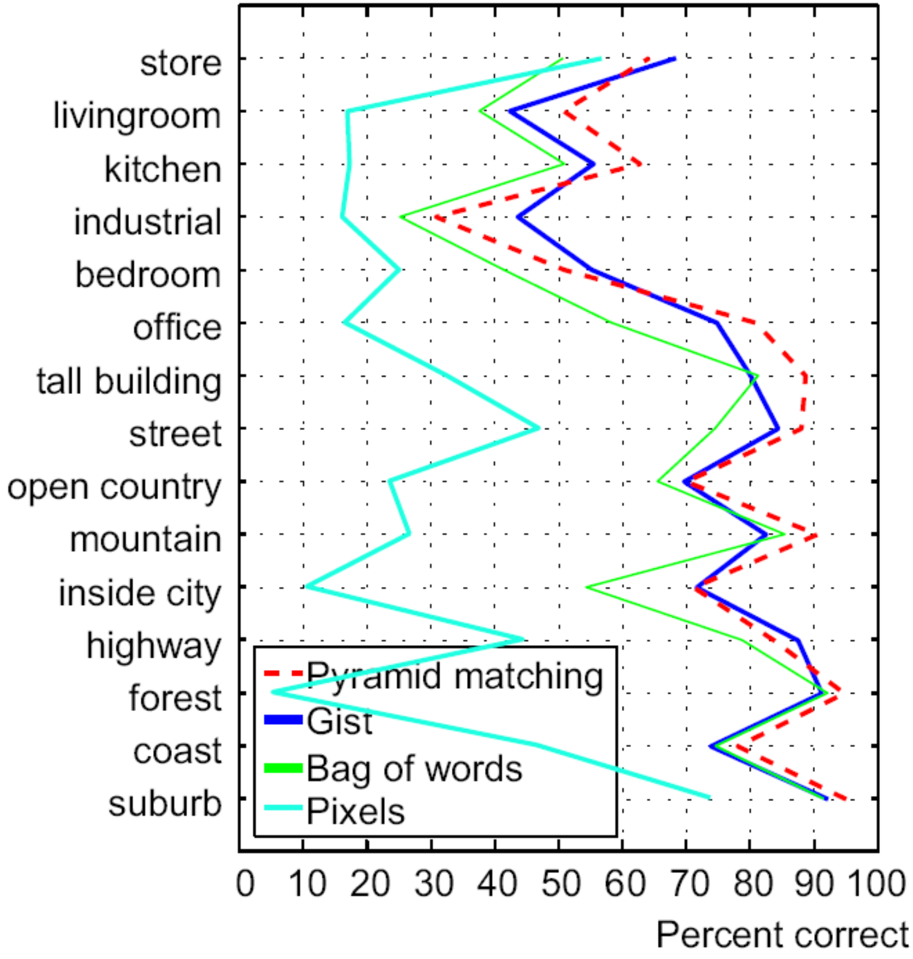
Forest path
Vs.
all

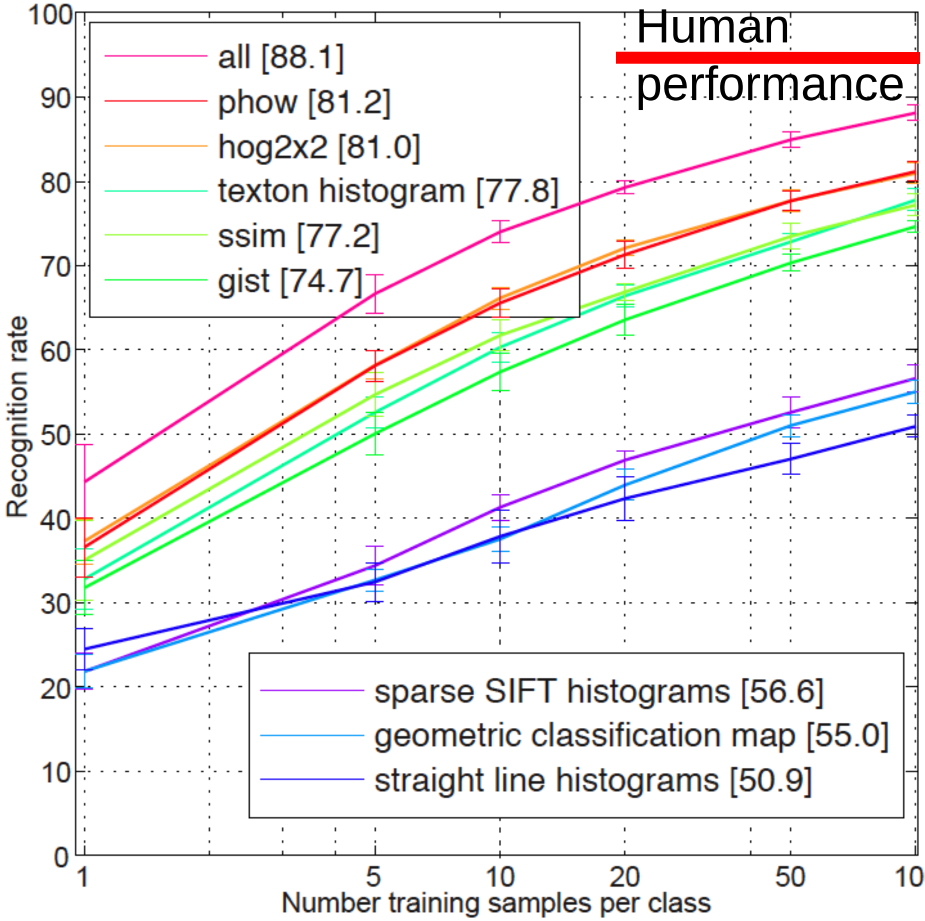
Living - room
Vs.
all

Scene recognition

100 training samples per class

SVM classifier in both cases

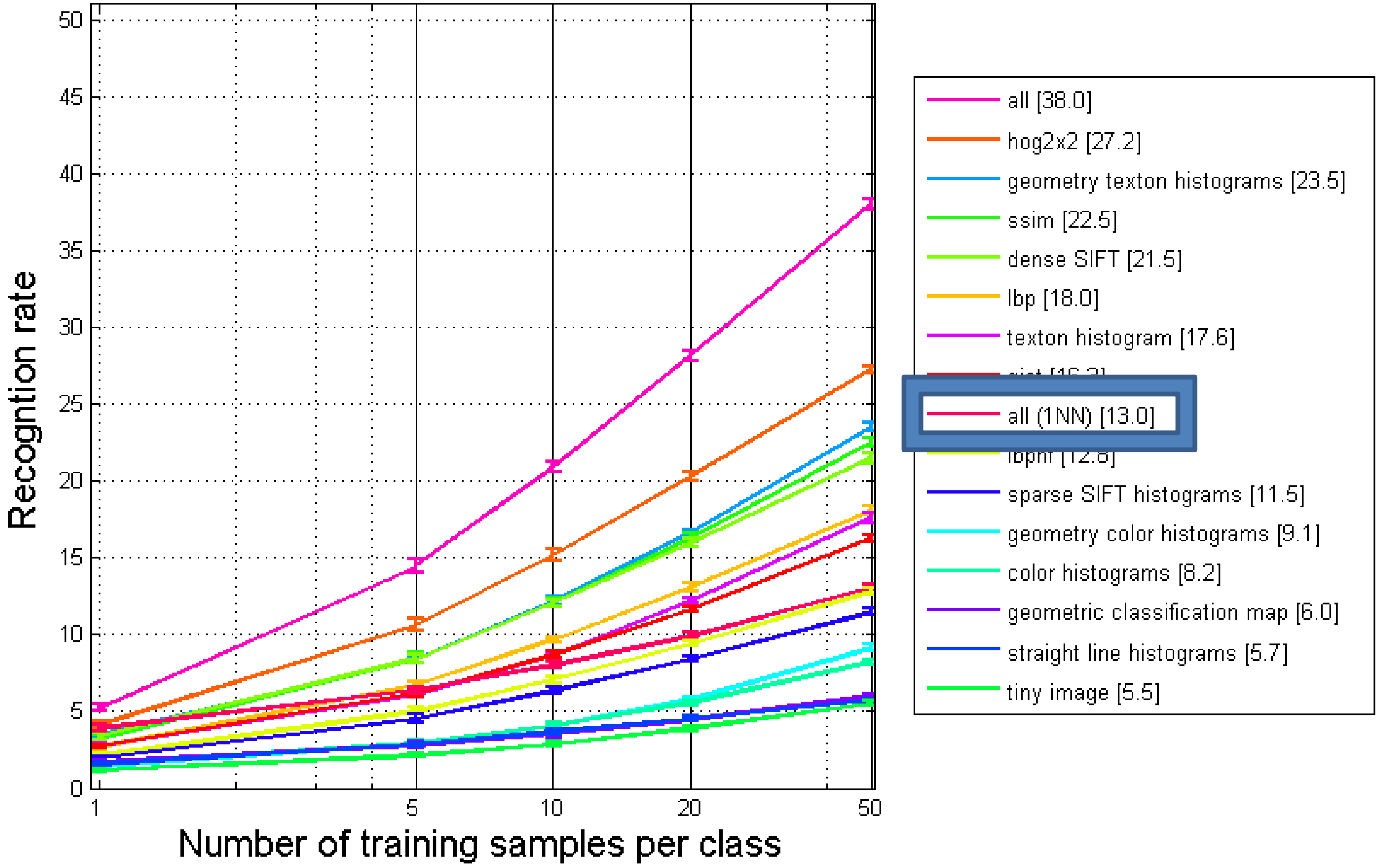




Human performance

Feature Accuracy

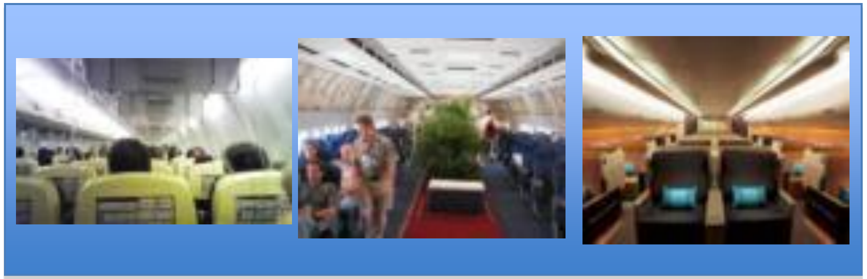
Humans [68.5]

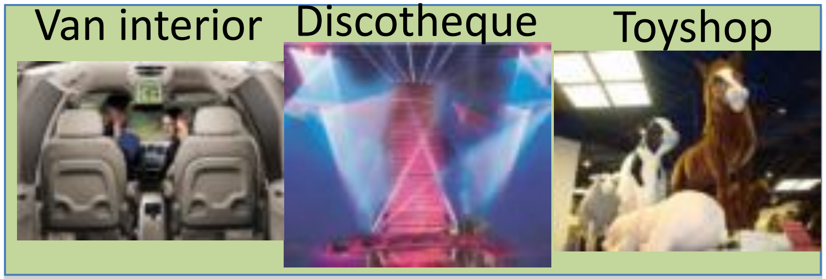


Classifier: 1-vs-all SVM with histogram intersection, chi squared, or RBF kernel.

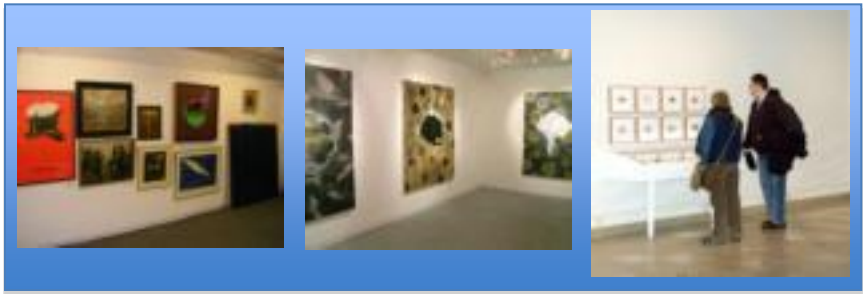
A look into the results

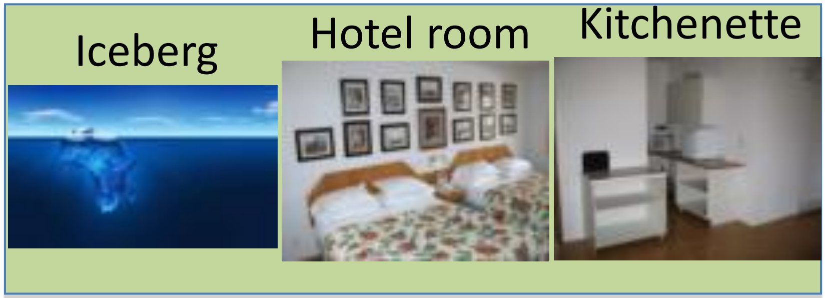
Airplane cabin (64%)





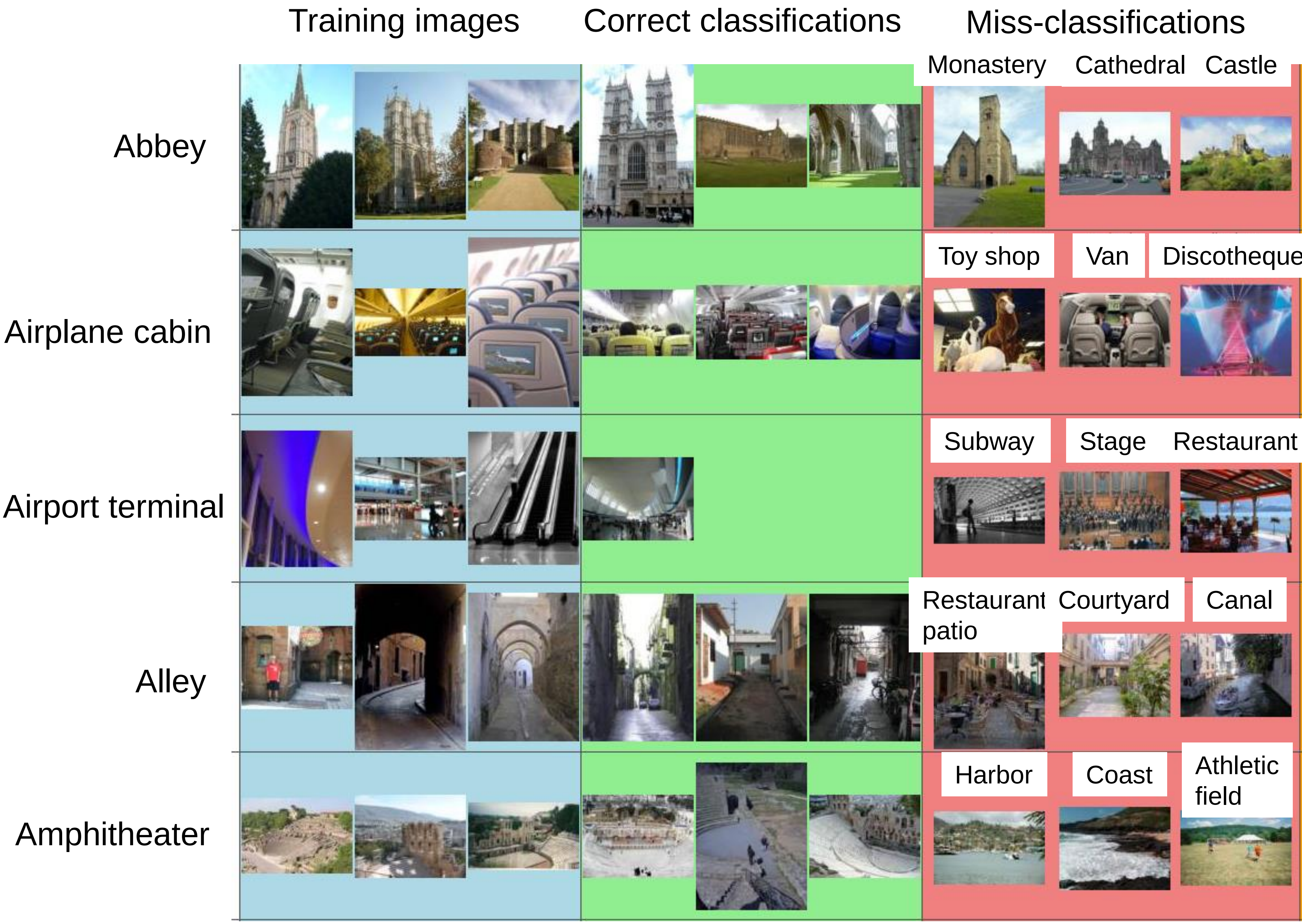
Art gallery (38%)





All the results available on the web

…



Xiao, Hays, Ehinger, Oliva, Torralba; maybe 2010

Humans good
Comp. good

limousine interior
(95% vs 80%)

riding arena
(100% vs 90%)

sauna
(96% vs 95%)

skatepark
(96% vs 90%)

subway interior
(96% vs 80%)

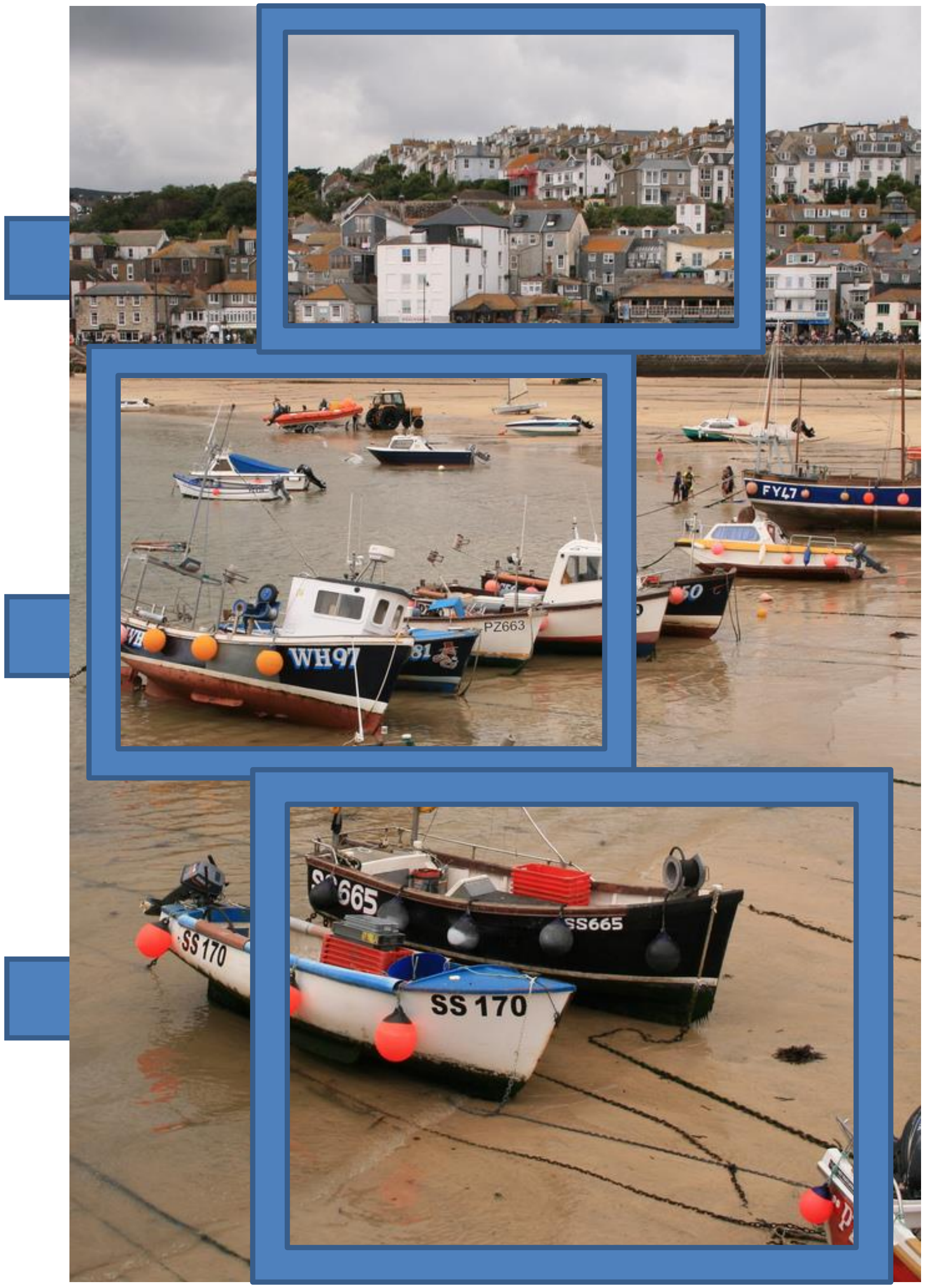
Humans bad
Comp. bad

Human good
Comp. bad

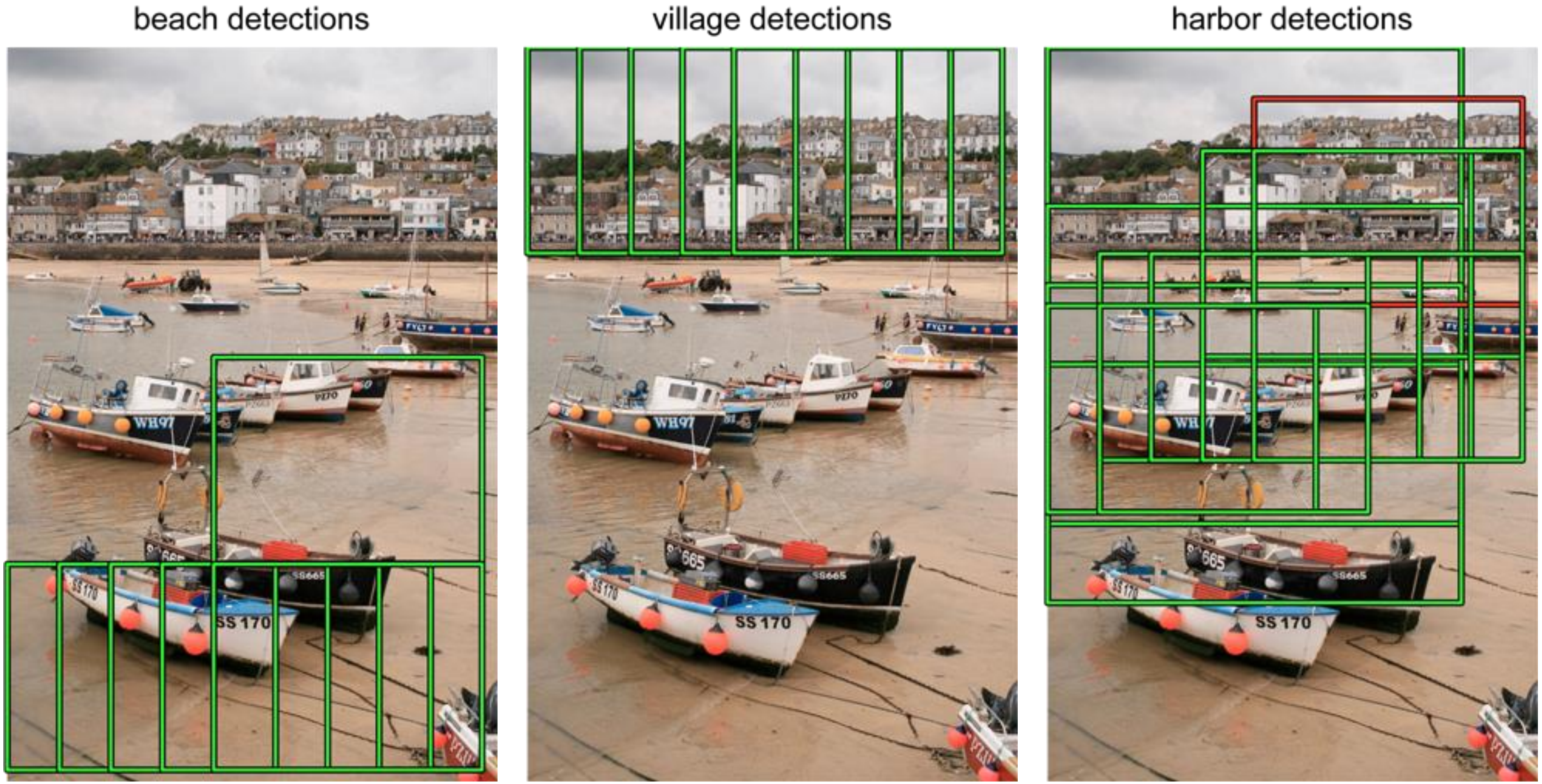
Human bad
Comp. good

SS665
SS 170

WH97
PZ663


FY47
WH97
PZ663
SS665
SS 170

Local Scene Detection

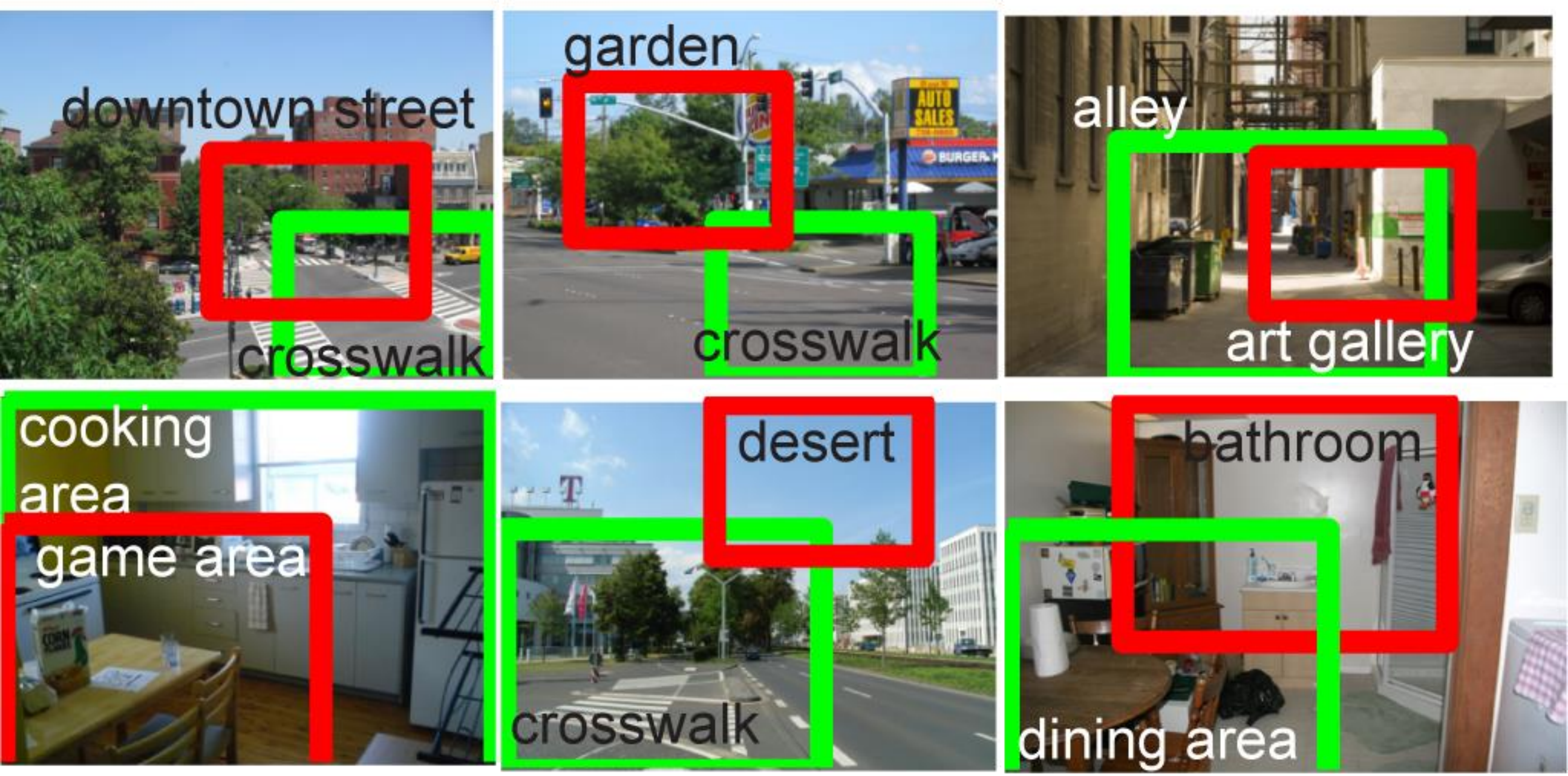


beach detections

village detections

harbor detections

Confident Subscene Detections



Database and source code available at

http://groups.csail.mit.edu/vision/SUN/

Additional details available:

SUN Database: Large-scale Scene Recognition from Abbey to Zoo. Jianxiong Xiao, James Hays, Krista A. Ehinger, Aude Oliva, Antonio Torralba. *CVPR 2010.*

How do we do better than 40%?

- Deep learning gets about the same performance
- Fisher vector encoding gets up to 47.2%