Scene representation I Generalities

D.A. Forsyth

High level issues

- What kind of representation should we make?
  - 3D, 2D, Biased, Unbiased,
- With what perceptual inputs?
- Analyzed how?

Structure

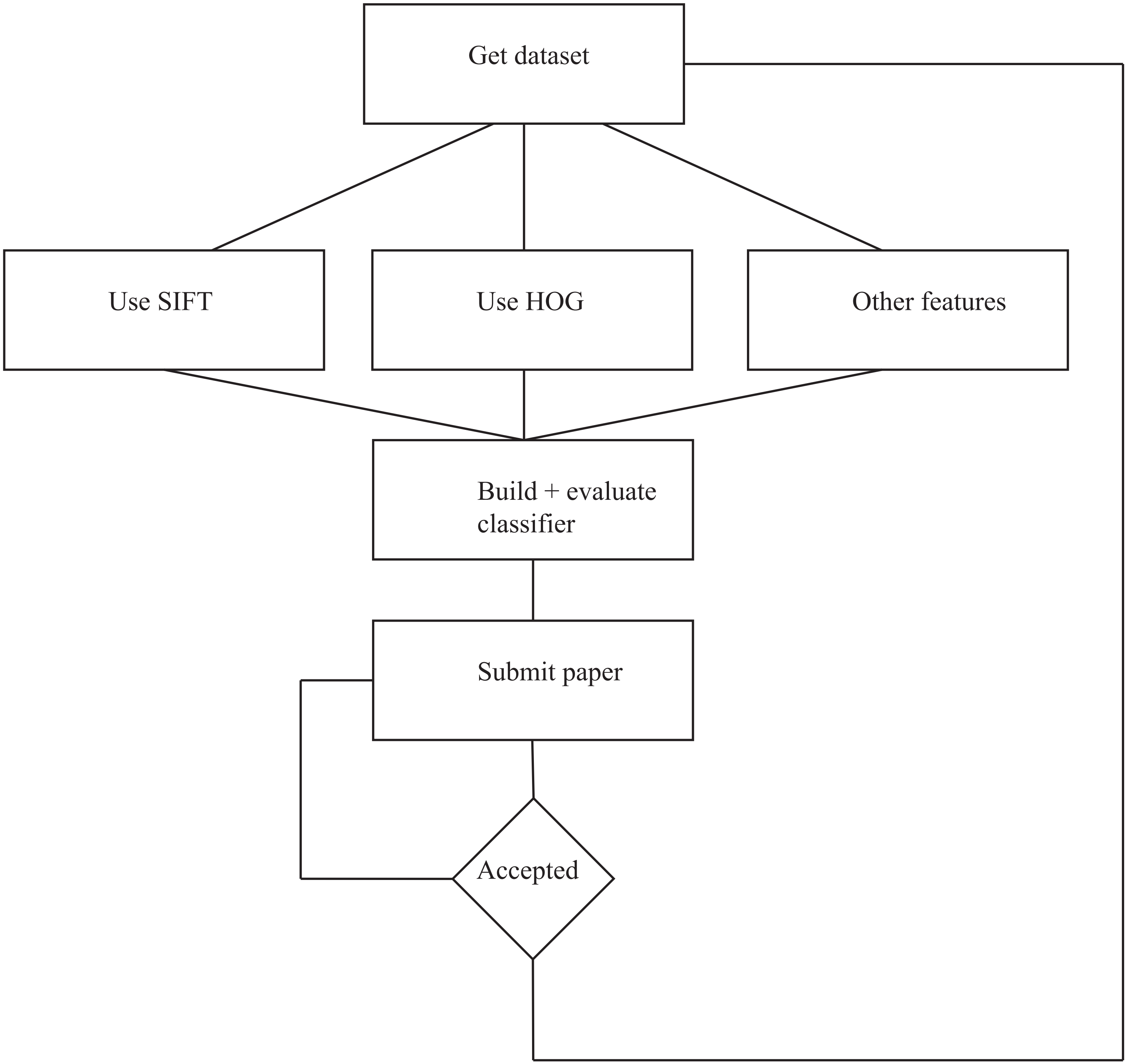
- Recognition has much more to do than object tagging
  - potential and scenes
- Indoor spaces, bias and variance
  - there is a bias-variance tradeoff in modeling that is still poorly understood
  - good models can be recovered automatically (or nearly)
    - from single images
    - from RGBD
  - such models can be used to reason about potential

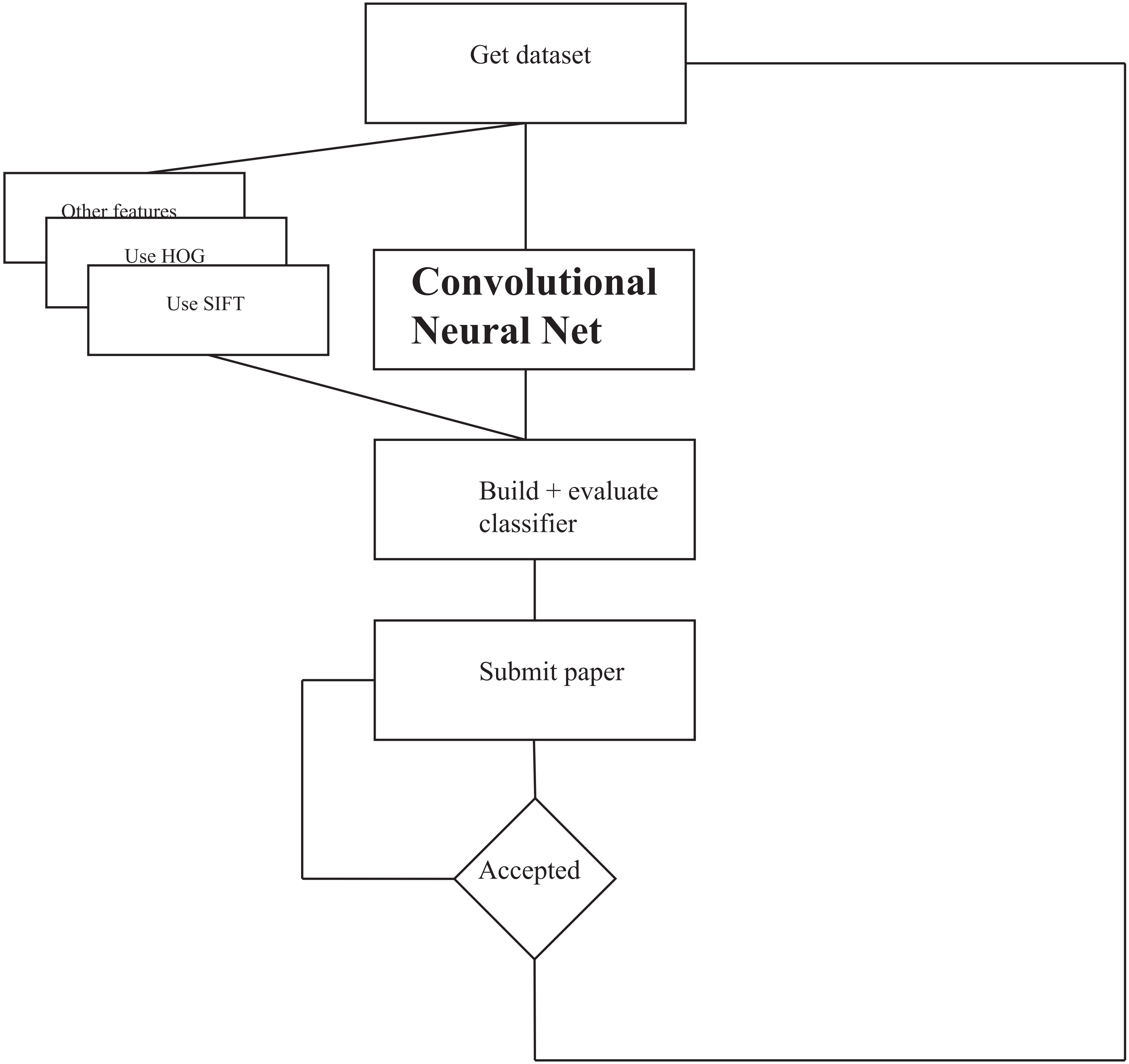
The idea of potential

Platonism?

A belief space about recognition

- Object categories are fixed and known
  - Each instance belongs to one category of k
- Good training data for categories is available
- Object recognition=k-way classification
- Detection = lots of classification


Get dataset
Use SIFT
Use HOG
Other features
Build + evaluate classifier
Submit paper
Accepted


Get dataset
Other features
Use HOG
Use SIFT
Convolutional Neural Net
Build + evaluate classifier
Submit paper
Accepted

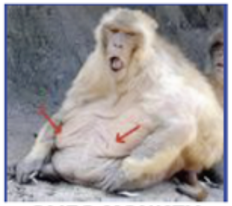
Platonism?

A belief space about recognition

- Object categories are fixed and known — Obvious nonsense
  - Each instance belongs to one category of k — Obvious nonsense
- Good training data for categories is available — Obvious nonsense
- Object recognition=k-way classification
- Detection = lots of classification

Are these monkeys?

Spider **Monkey**, Spider **Monkey** Profile ...
470 x 324 - 29k - jpg
animals.nationalgeographic.com
[More from animals.nationalgeographic.com]



OMFG **MONKEY** NIPS2.
444 x 398 - 40k - jpg
www.bestweekever.tv
[More from www.bestweekever.tv]

Vampire **Monkey**
350 x 500 - 32k - jpg
paranormal.about.com

... **monkeys** for ...
424 x 305 - 21k - jpg
thebitt.com

The **Monkey** Cage
300 x 306 - 35k - jpg
www.themonkeycage.org

... be **monkey** ...
300 x 350 - 29k - jpg
my.opera.com

... **monkey's** interests ...
378 x 470 - 85k - jpg
www.schwimmerlegal.com

"You will be a **monkey**.
358 x 480 - 38k - jpg
kulxp.blogspot.com

... **monkey** and I am ...
342 x 324 - 17k - jpg
www.azcazandco.com

Monkey
353 x 408 - 423k - bmp
www.graphicshunt.com

The **Monkey** Park
400 x 402 - 24k - jpg
www.lysator.liu.se

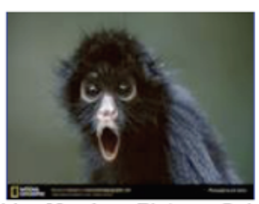
Monkey cloning follow up ...
450 x 316 - 17k - jpg
blog.bioethics.net

So here's one of my **monkeys**.
400 x 300 - 13k - jpg
www.gamespot.com

monkeys ...
400 x 310 - 85k - jpg
joaquinvargas.com

MONKEY TEETH
308 x 311 - 18k - jpg
repairstemcell.wordpress.com

The Blow **Monkey** is ...
500 x 500 - 30k - jpg
www.uberreview.com

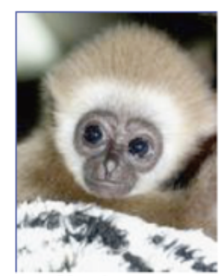


Spider **Monkey** Picture, Spider **Monkey** ...
800 x 600 - 75k - jpg
animals.nationalgeographic.com

a........ **monkey**! mammal **monkey**
525 x 525 - 99k - jpg
www.sodahead.com

WTF **Monkey**
374 x 300 - 23k - jpg
www.myspace.com

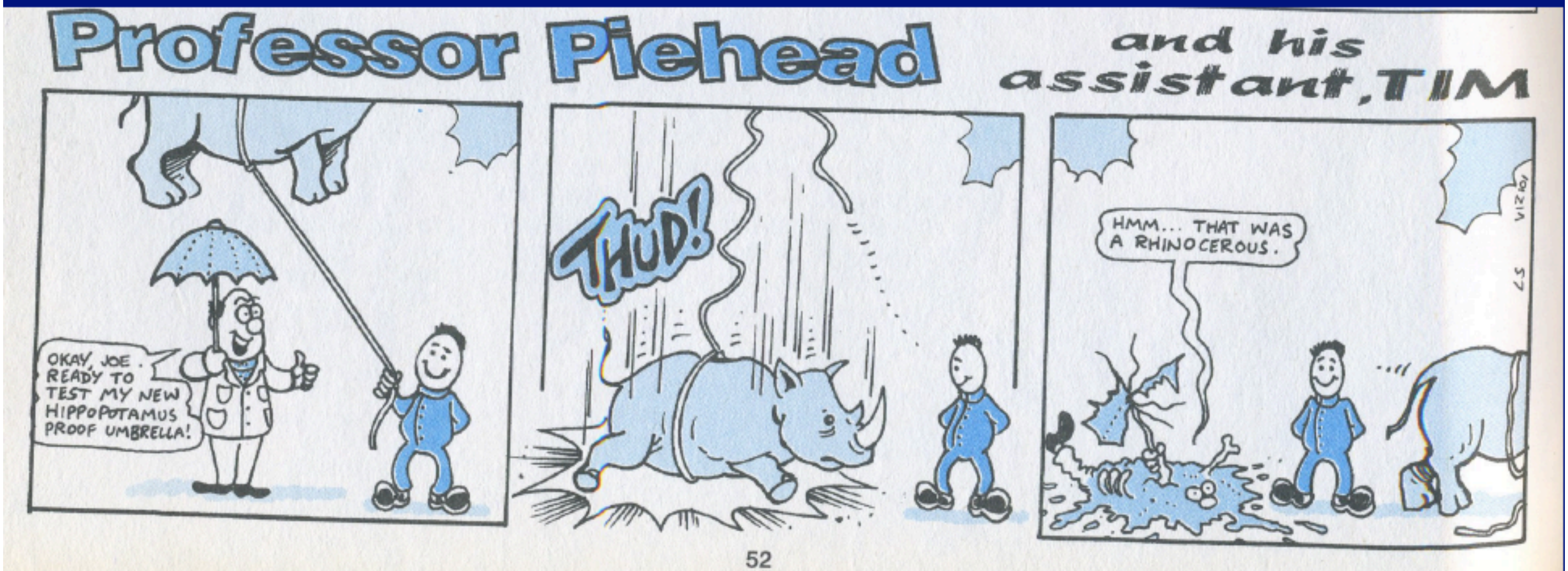
Monkey
512 x 768 - 344k - jpg
www.exzooberance.com



Monkeys ...
787 x 1024 - 131k - jpg
runrigging.blogspot.com

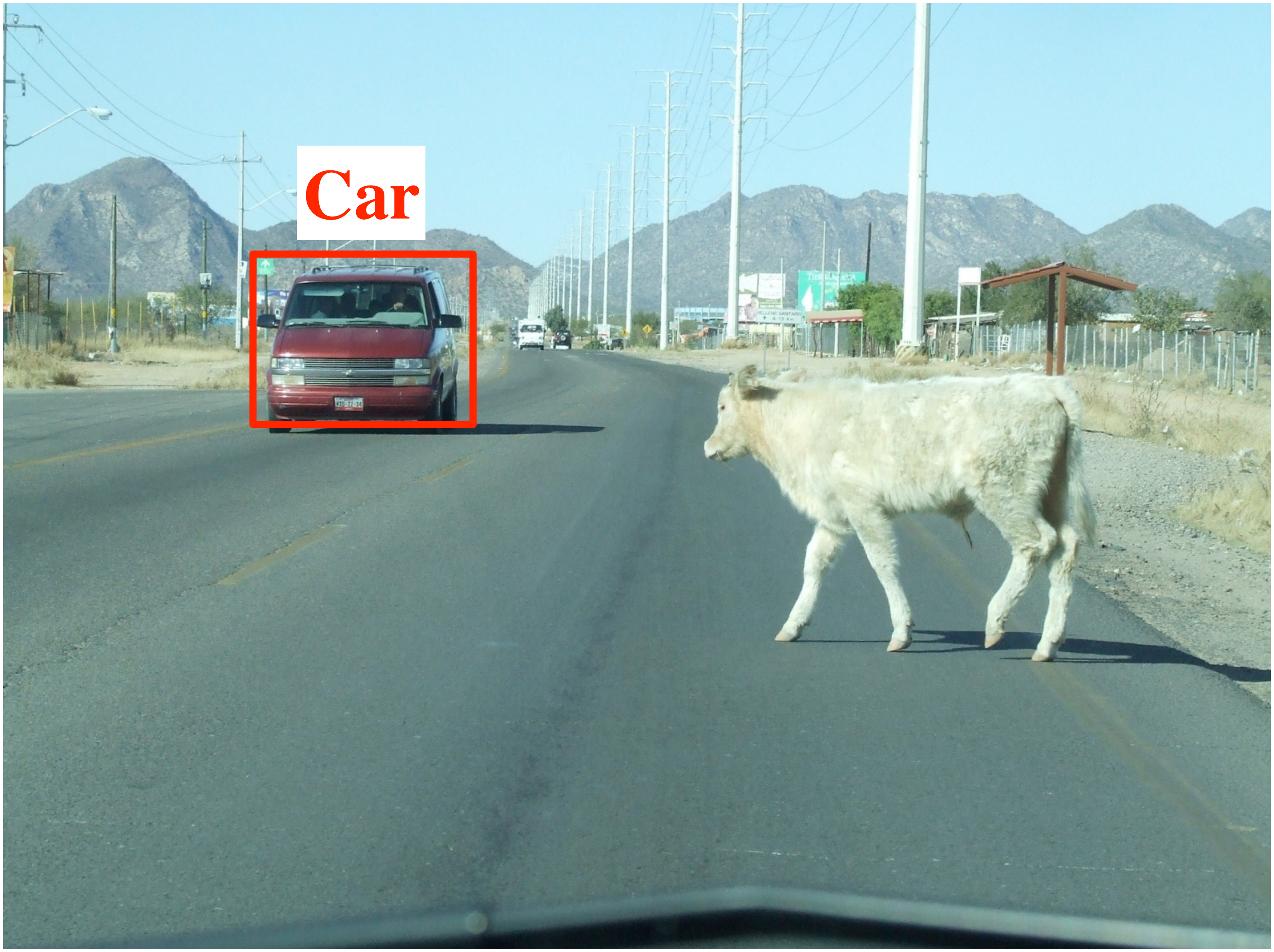
What have we inherited from this view?

- Deep pool of information about feature constructions
- Tremendous skill and experience in building classifiers
- Much practice at empiricism
  - which is valuable, and hard to do right



Viz comic, issue 101

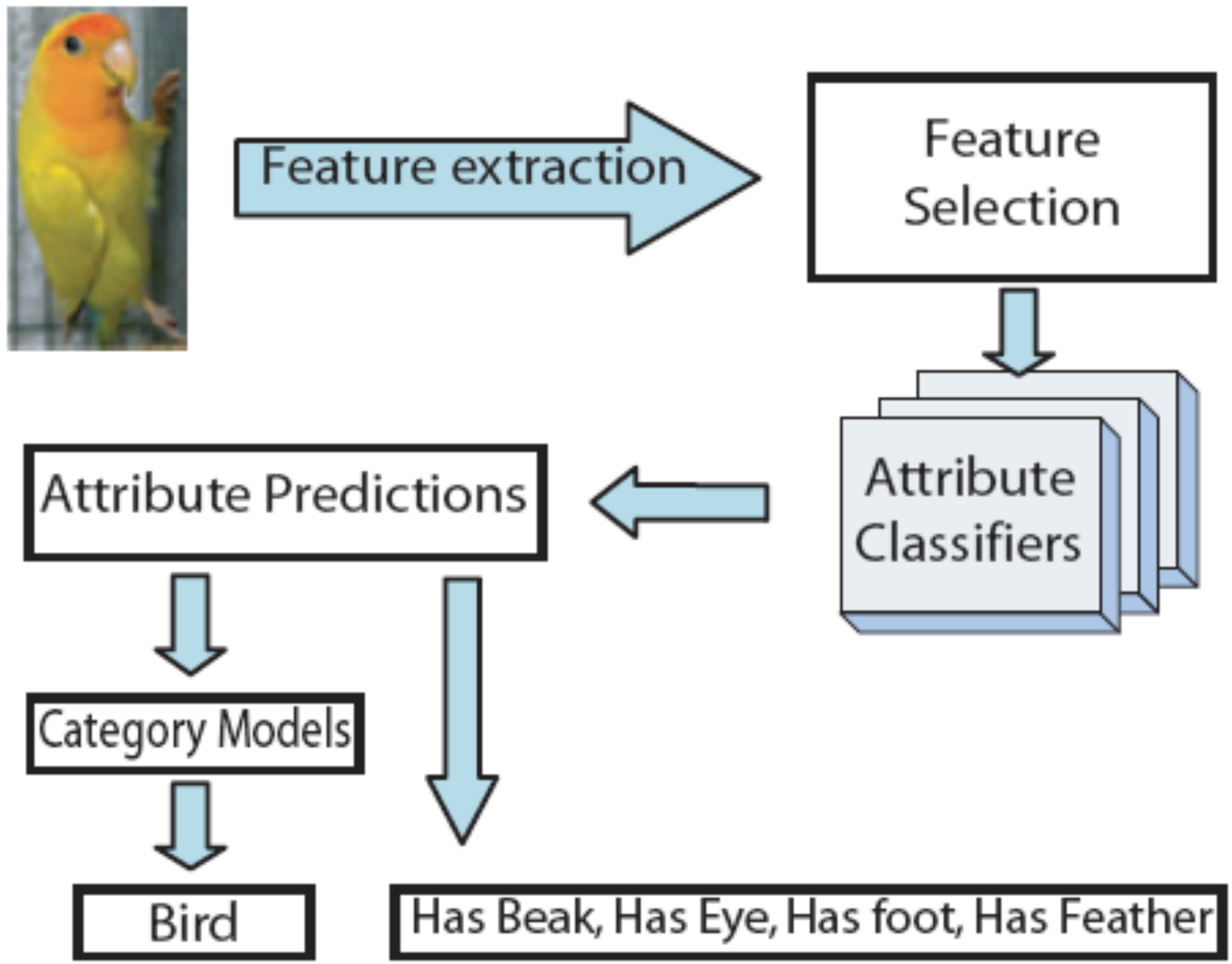
Coping with the unfamiliar


Car

Current strategies for coping

- Attributes
  - describe things by properties
  - a small “vocabulary” describes many different objects
- Affordances
  - geometric properties that expose “what an object is for”
  - a small “vocabulary” describes many different objects
- Primitives
  - a small “vocabulary” makes up many different objects
  - typically, shapes, but that isn’t compulsory
    - eg shared parts; texture encodings; deep learning

Attributes



Farhadi et al 09; cf Lampert et al 09

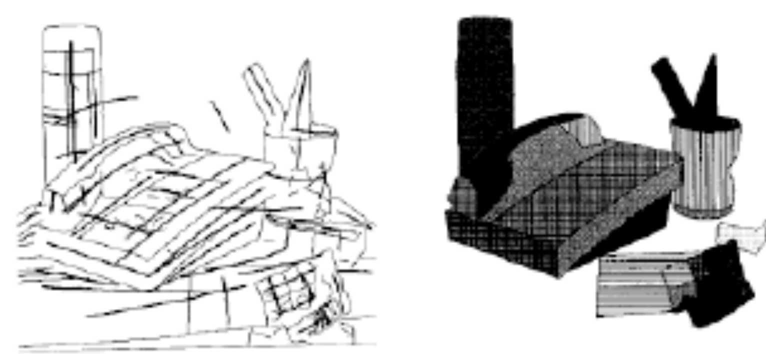
Attribute predictions for unknown objects

Farhadi et al 09; cf Lampert et al 09

Primitives allow joining up evidence

- Because only some patterns are possible
  - eg
    - everything's a generalized cylinder
    - => edges can only make objects in particular ways
    - => parse into generalized cylinders

Edges   Joined curves   Symmetry axes



Best symmetry axes   Surface patches

Mohan and Nevatia, 1989

The problem

- What primitives/attributes/affordances describe the world?
- How do you learn which ones describe the world?
- How do you ensure that the vocabulary is small
  - even if the set of objects is large?

What does vision do?

- Lists object names (?)
- Lists object descriptions (?)

- Evokes emotional states
  - but what do we do about this?

- Exposes possible futures
  - What could happen
  - Where you could go
  - Who could move close to you
  - What could be useful for

We should think about potential, rather than just or as well as, actual

Nobody was hurt in the coming movie


10/2009
PM

How many adults were on the platform and what were they doing?

How many benches were on the platform?

Were there flowers on the platform?

Was there a “no smoking” sign?

What outcome do we expect?

How are other people feeling?

What will they do?

What’s going to happen to the baby?


10/2009
PM

What outcome do we expect?

How are other people feeling?

What will they do?

What’s going to happen to the baby?

How many adults were on the platform and what were they doing?

How many benches were on the platform?

Were there flowers on the platform?

Was there a “no smoking” sign?

What outcome do we expect?

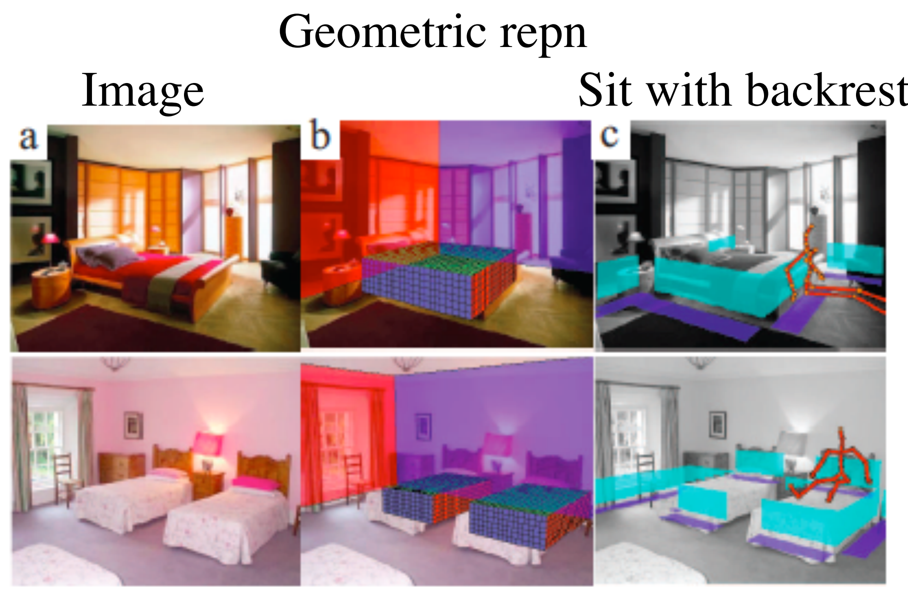
How are other people feeling?

What will they do?

RapidABC data

Potential

- What could
  - I do; happen to me; occur in the world
- Free space has motion potential
  - I could move there; things could move there to me; etc
- Free space has light potential
  - light goes through it
- Objects have potential
  - they can do things; or be done to; or be done with; etc.
- People have potential
  - what next?



The idea of a scene

Definition

- A scene is a view of a **real-world environment** that contains **multiples** surfaces and objects, organized in a **meaningful way**.

- Distinction between objects and scenes:

objects are compact and act upon

Scenes are extended in space and act within

The distinction depends on the action of the agent

Oliva slides

A few facts about human scene understanding

- Immediate recognition of the *meaning* of the scene and the *global structure*

- Quick visual perception lacks of objects and details information. Objects are *inferred, not necessarily seen*

Oliva slides

Which One Did You See?

A

B

C

D

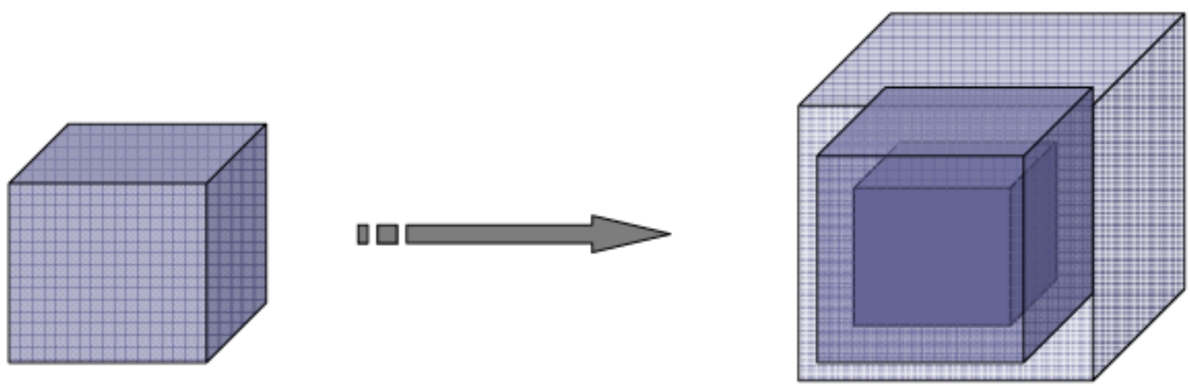
Systematic scene memory *distortion*

correct answer

B **C**

too close ⟷ **too far**

You tend to remember seeing more of a scene than was there



Helene Intraub (Boundary Expansion Effect on pictures of object)

The Gist of the Scene

- Mary Potter (1975, 1976) demonstrated that during a rapid sequential visual presentation (100 msec per image), a novel scene picture is indeed instantly **understood** and observers seem to comprehend a lot of visual information, **but a delay of a few hundreds msec (~ 300 msec) is required for the picture to be consolidated in memory.**

- The "gist" (a summary) refers to the visual information perceived after/during a glance at an image.

- To simplify, the gist is often synonymous with the *basic-level category* of the scene or event (e.g. wedding, bathroom, beach, forest, street)

What is represented in the gist ?

- The "Gist" includes all levels of visual information, from low-level features (e.g. color, luminance, contours), to intermediate (e.g. shapes, parts, textured regions) and high-level information (e.g. semantic category, activation of semantic knowledge, function)

- **Conceptual gist** refers to the semantic information that is inferred while viewing a scene or shortly after the scene has disappeared from view.

- **Perceptual gist** refers to the structural representation of a scene built during perception (~ 200-300 msec).

Oliva, A. (2005). Gist of a scene. In *Neurobiology of Attention*. Eds. L. Itti, G. Rees and J. Tsotsos. Academic Press, Elsevier.

Some simple features are correlated with scene recognition

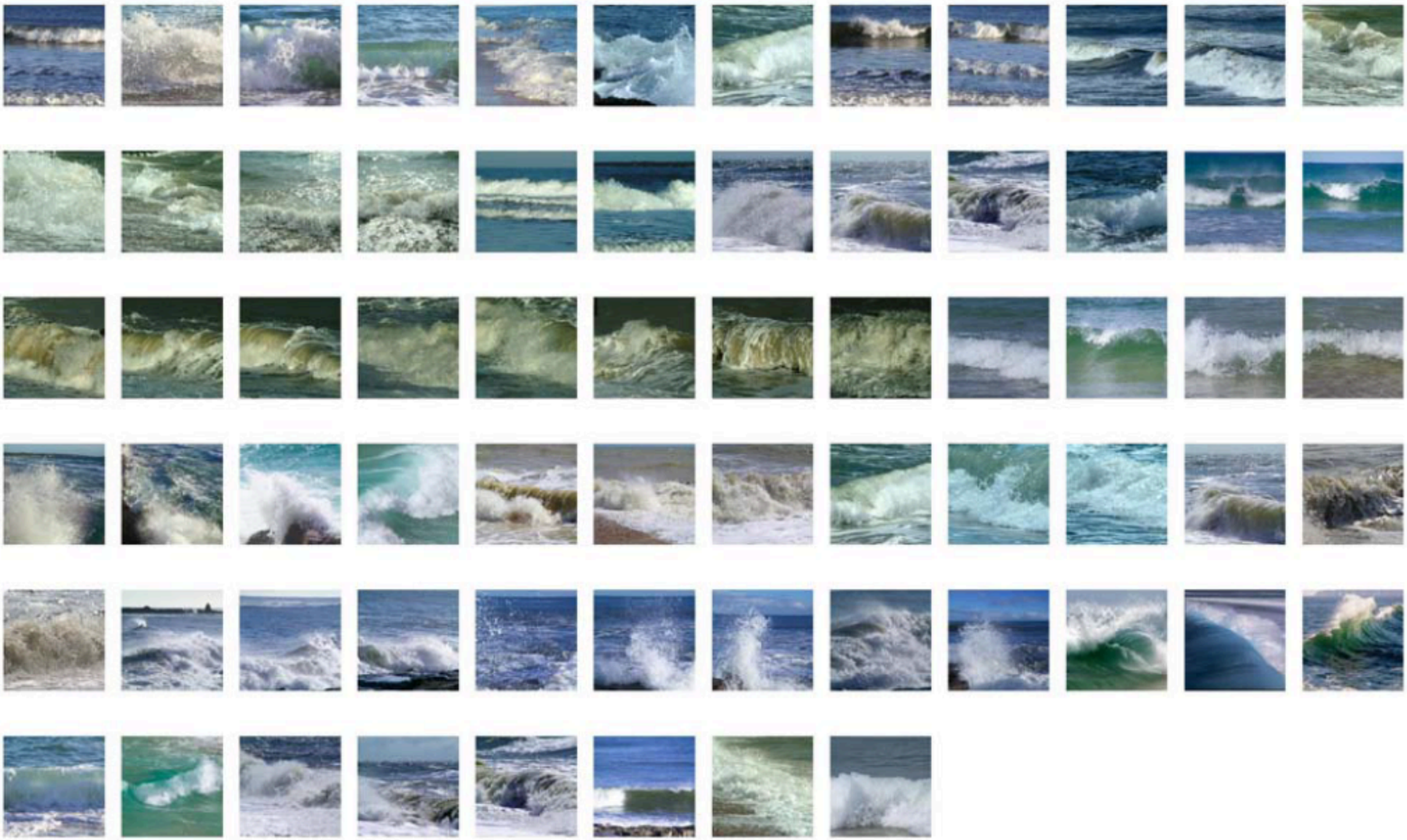
What are the other properties of a scene image that could help “recognition” (gist)?

Navon (1977) says:

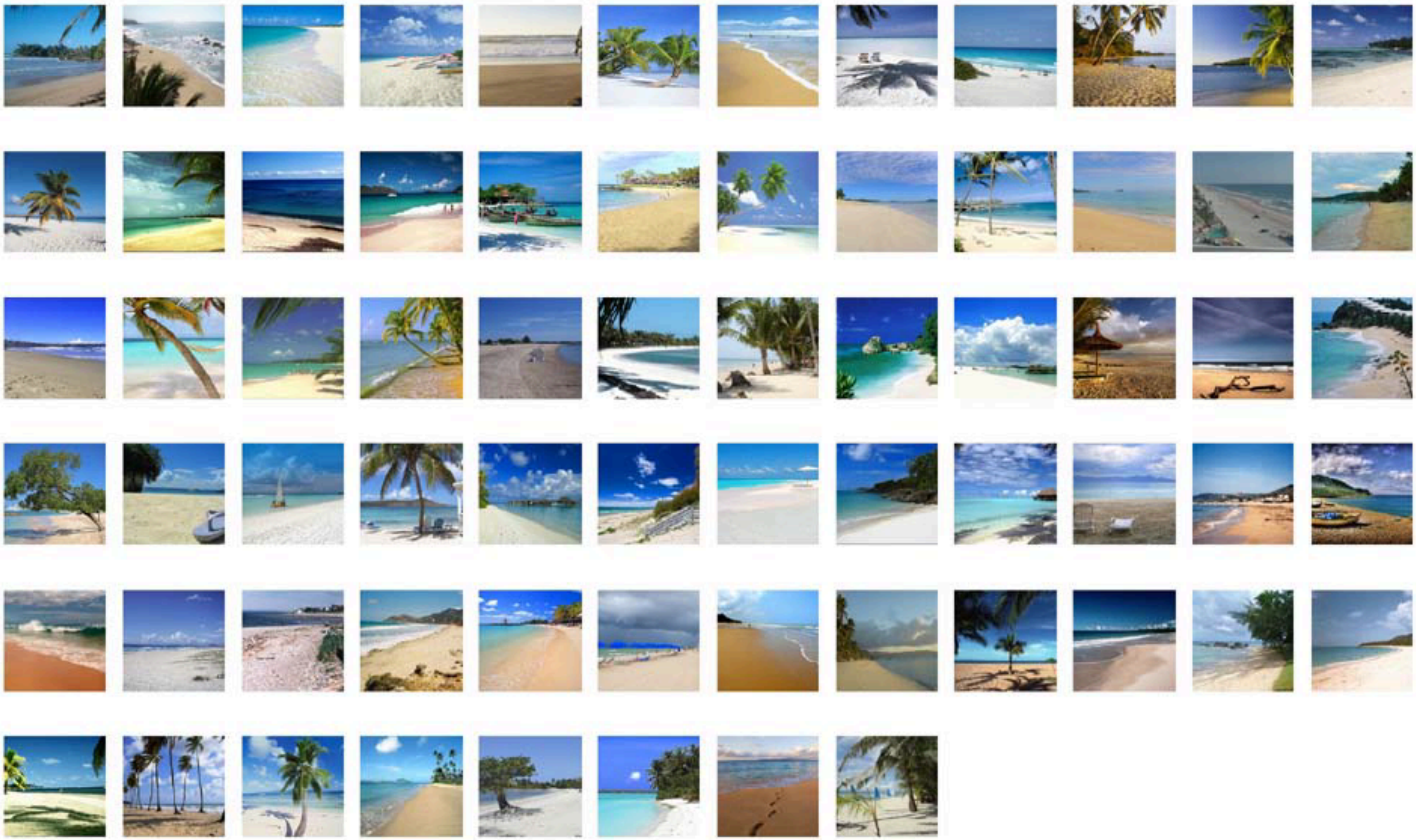
- “No attempt was made here to formulate an operational definition of globality of visual features which enables precise predictions about yhe course of perception of real-world scenes.
- What is suggested in this paper is that whatever the perceptual units are, the spatial relationship among them is more global than the structure within them (and so forth if the hierarchy is deeper).
- Thus, I am afraid that clear-cut operational measures for *globality* will have to patiently await the time that we have a better idea of **how a scene is decomposed into perceptual units. “**

What are perceptual units?

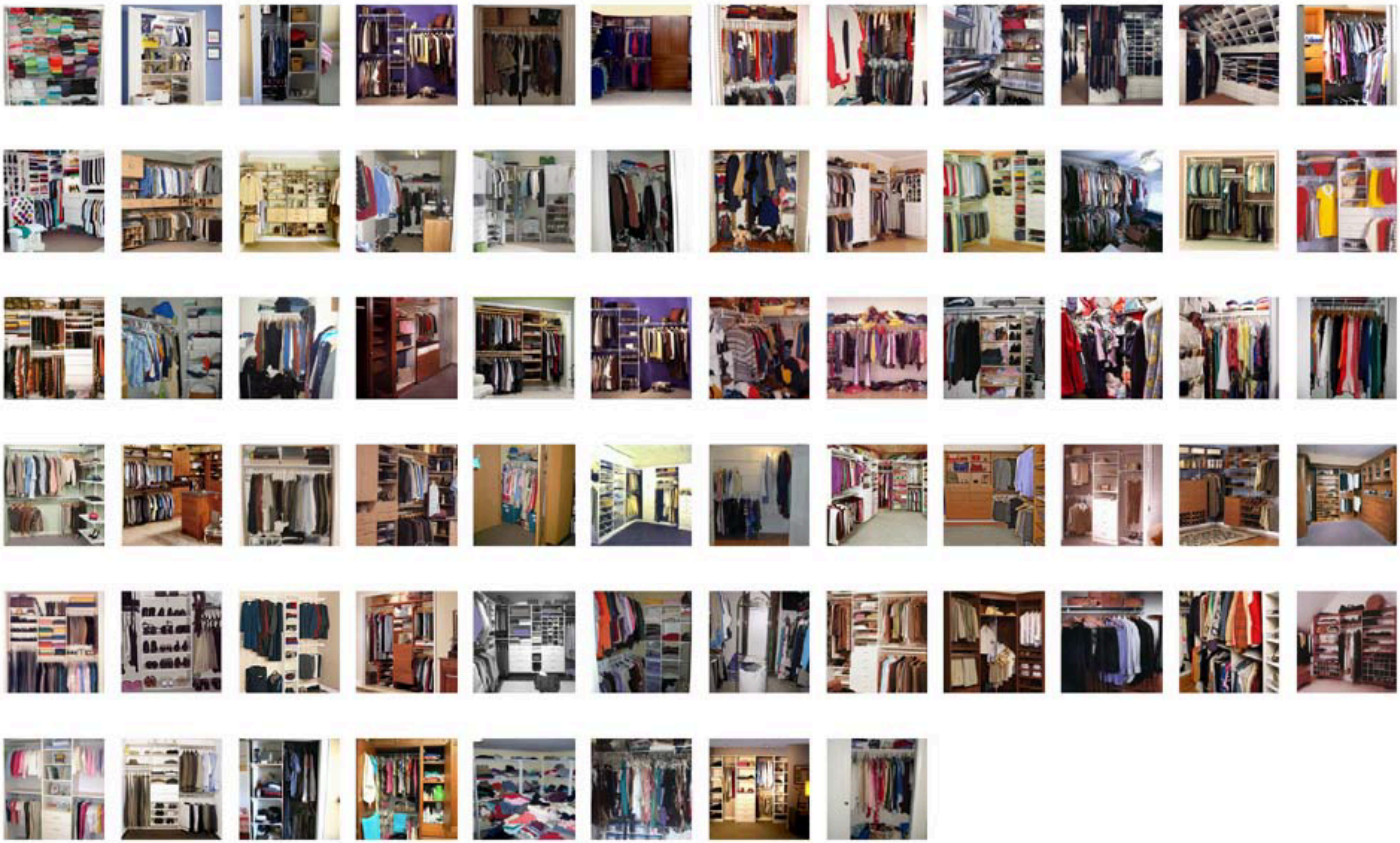
Waves ~ Texture



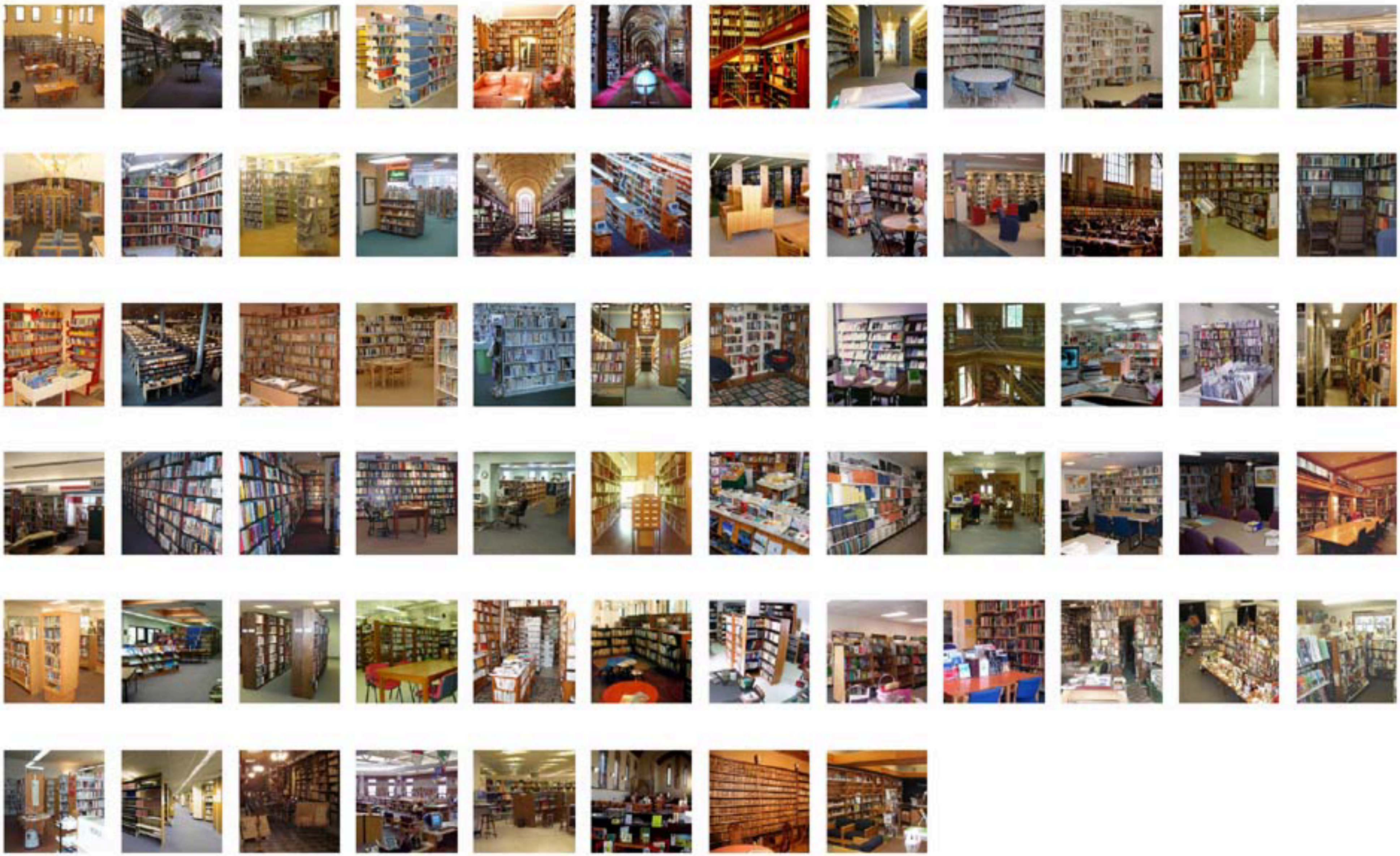
Beach



Closet



Library



Part-based approach: e.g. *objects*

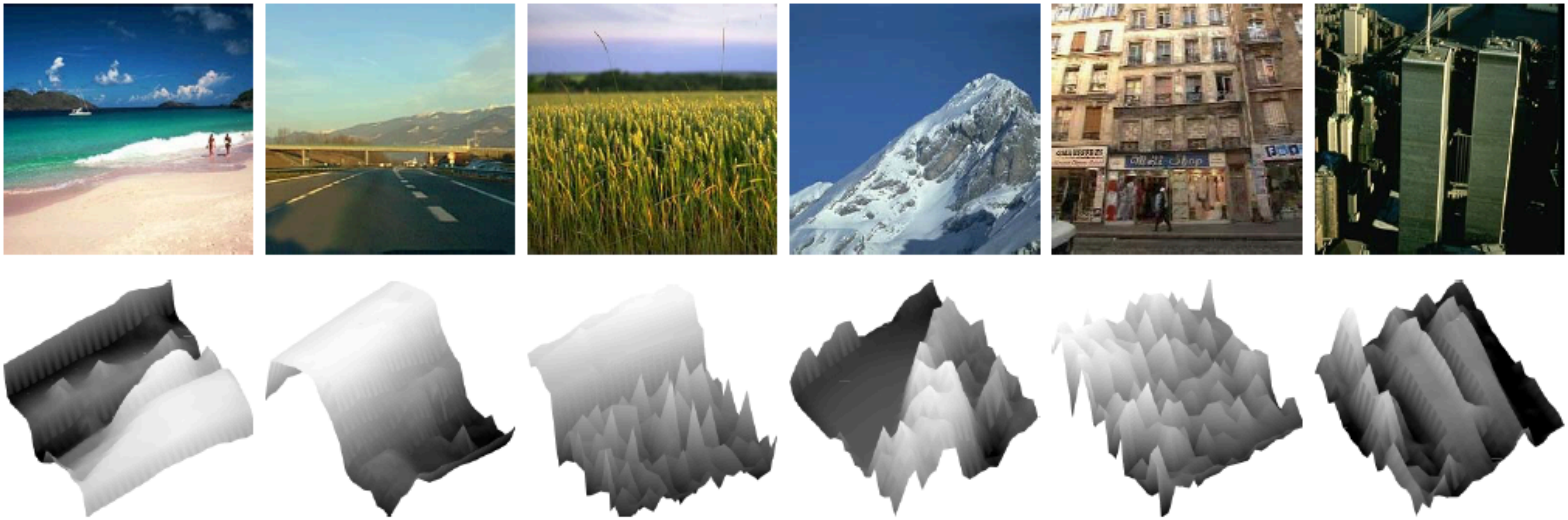
If you knew the identity of all the objects in a scene, recognition would be perfect

	Bathroom	Bedroom	Conf. room	Dining room	Kitchen	Living room	Office
Bathroom	99				03		
Bedroom		99	02				03
Conference room		02	98	01			07
Dining room			01	98	02	03	03
Kitchen	03			02	99		
Living room				03		99	01
Office		03	07	03		01	97

Labelme: a vector of the list of all objects for each image

Oliva et al. 2006

Holistic approach: global surface properties

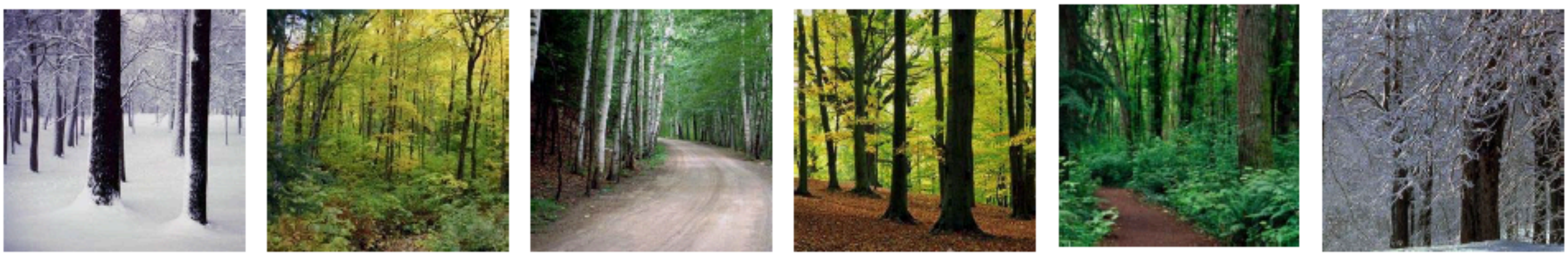


A scene is a single surface that can be represented by global descriptors

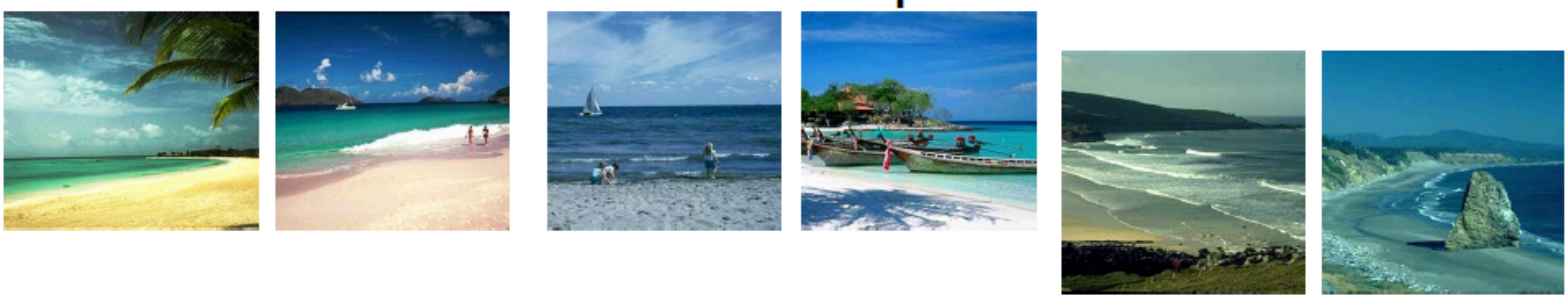
Oliva & Torralba (2001)

Hints of Globality: Spatial Structure

Forests are "enclosed"



Beaches are "open"



“Agnosic” human scene representation: How far can we go with it ?

A lake

Scene-Centered Representation

100% natural space
66% open space
64% perspective
74% deep space
68% cold place

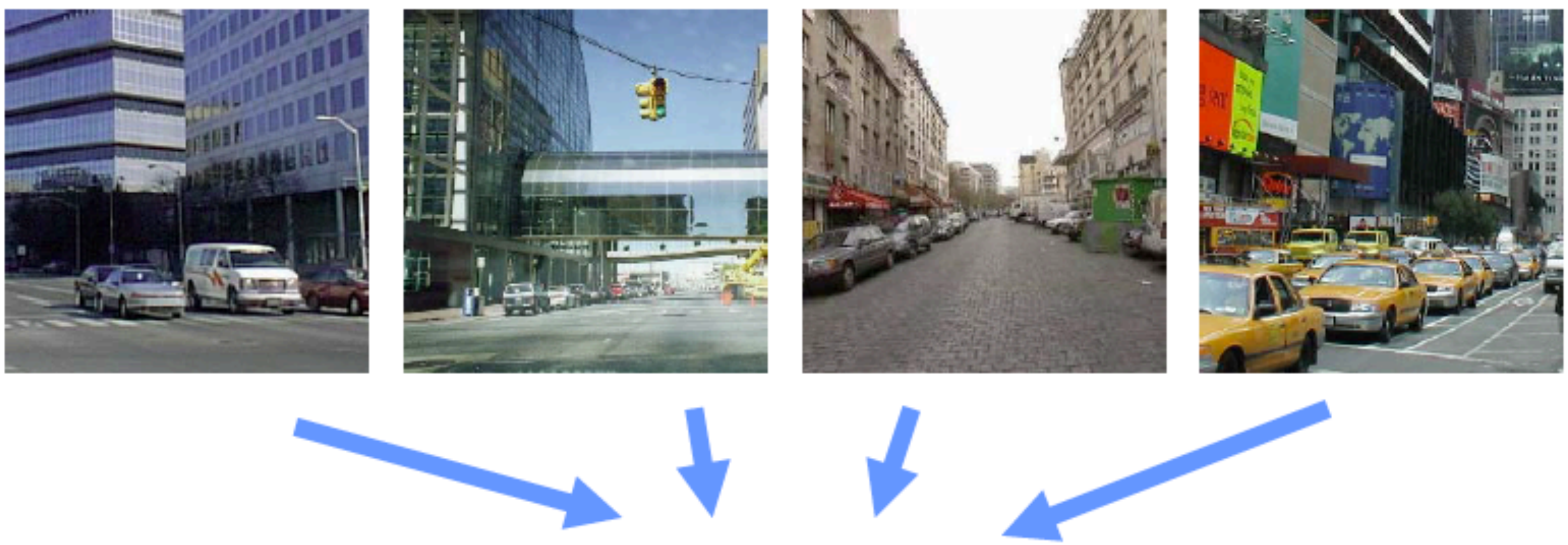
O ect-Cente d
esentatio

23 sky
35 % ter
18% tre
2 % mount
% gra

Spatial Envelope Theory

As a scene is inherently a 3D entity, initial scene recognition might be based on properties *diagnostic of the space* that the scene subtends and not necessarily the objects the scene contains

"Street"



Degree of clutter, openness, perspective, roughness, etc …

What is important for us here

- Early scene recognition methods
  - strongly emphasize “global shape” (GIST features, Oliva+Torralba 01)
  - effective, comparable to humans
- Recent methods
  - large scale classification (datasets in slides below)
  - no underlying feature theory
- Why do we care?
  - Our scenes have very stylized geometry
  - We should be able to benefit from this

Bias-Variance tradeoff

Best model in family

$$\mathbb{E}\left[(y-\hat{f}(x))^2\right] = \mathbb{E}\left[(y-f)^2\right] + \mathbb{E}\left[(f-\mathbb{E}\left[\hat{f}\right])^2\right] + \mathbb{E}\left[(\mathbb{E}\left[\hat{f}\right]-\hat{f})^2\right]$$

Chosen model

- Expected error in predictions consists of three terms
  - easily proved (look it up; do it yourself)
  - expectation taken over all possible choices of training data

Bias-Variance tradeoff

$$\mathbb{E}\left[(y-\hat{f}(x))^2\right] = \underbrace{\mathbb{E}\left[(y-f)^2\right]}_{\text{Error resulting from choice of family}} + \underbrace{\mathbb{E}\left[(f-\mathbb{E}\left[\hat{f}\right])^2\right]}_{\text{Error resulting from BIAS of learning algorithm}} + \underbrace{\mathbb{E}\left[(\mathbb{E}\left[\hat{f}\right]-\hat{f})^2\right]}_{\text{Error resulting from VARIANCE of learning algorithm}}$$

These are affected by choice of model AND of algorithm

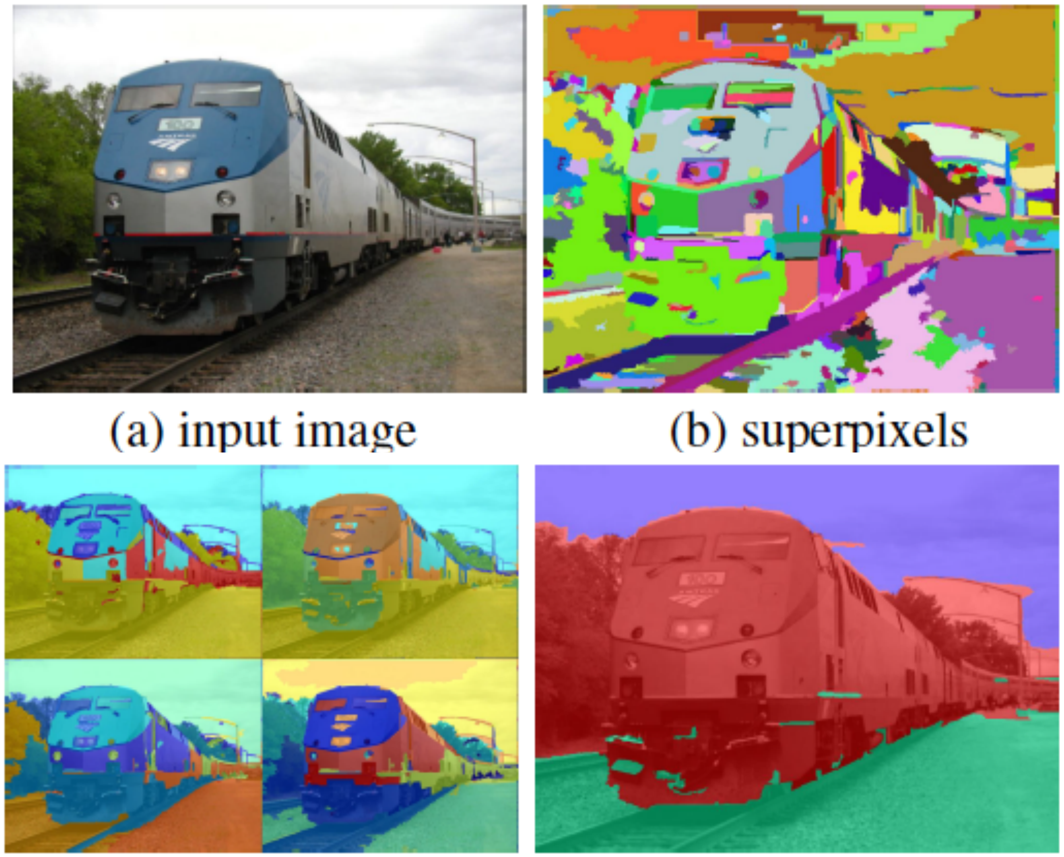
Bias-Variance tradeoff

$$\mathbb{E}\left[(y - \hat{f}(x))^2\right] = \mathbb{E}\left[(y - f)^2\right] + \mathbb{E}\left[(f - \mathbb{E}\left[\hat{f}\right])^2\right] + \mathbb{E}\left[(\mathbb{E}\left[\hat{f}\right] - \hat{f})^2\right]$$

Model Bias Learning Bias Variance

- Generally, these error terms trade off against one another
  - if one goes down, another goes up
  - because if the representation/algorithm are unbiased
    - you usually have to estimate MORE STUFF (and so make more errors)
- Variance is scary
  - bias, tends not to be
- Managing relationship is key in choosing representations

Photo Pop-up


(a) input image (b) superpixels

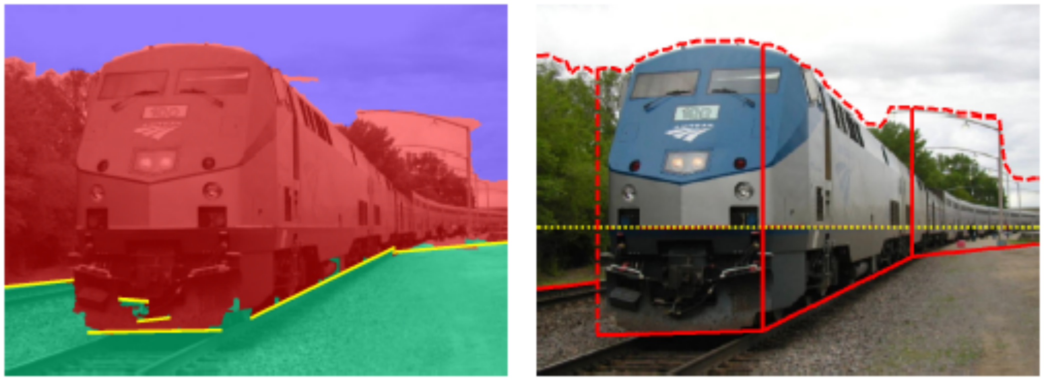
(c) constellations (d) labeling

Variance - method can't get these normals right

or even all these (though they're biased)

Hoiem et al 05

New view requires polygons

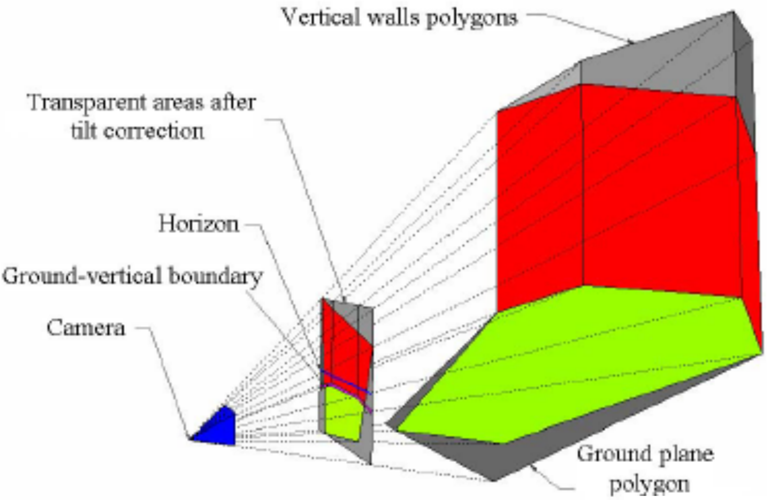


(a) Fitted Segments (b) Cuts and Folds

Figure 4: From the noisy geometric labels, we fit line segments to the ground-vertical label boundary (a) and form those segments into a set of polylines. We then “fold” (red solid) the image along the polylines and “cut” (red dashed) upward at the endpoints of the polylines and at ground-sky and vertical-sky boundaries (b). The polyline fit and the estimated horizon position (yellow dotted) are sufficient to “pop-up” the image into a simple 3D model.

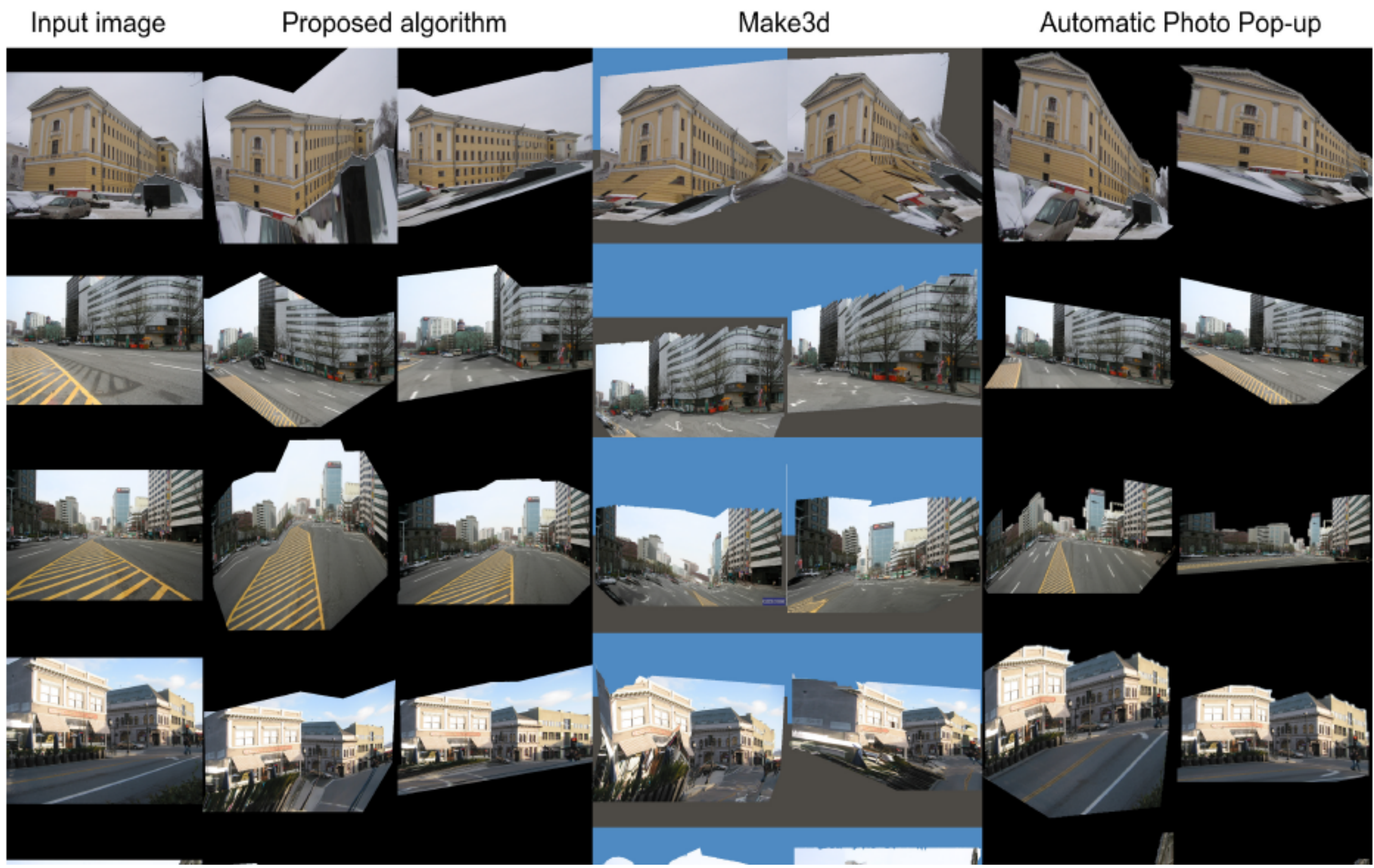
Hoiem et al 05

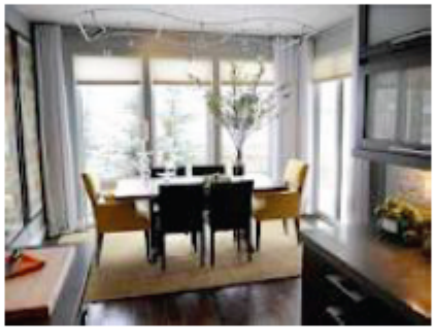
(e) novel view



More polygon representations

Barinova et al 08

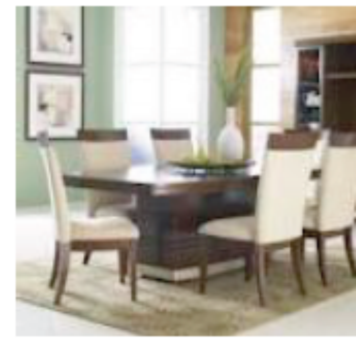
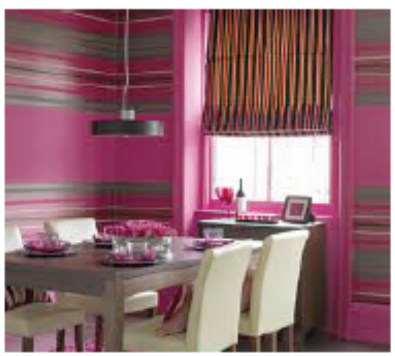












Fitting boxes to pics

Hedau et al 09

Comparison

Input Image | Hoiem *et al*. [9] | Liu *et al*. [14] | Barinova *et al*. [1] | Our Algorithm

Clutter does not need to be labelled

- Latent variables encode clutter points

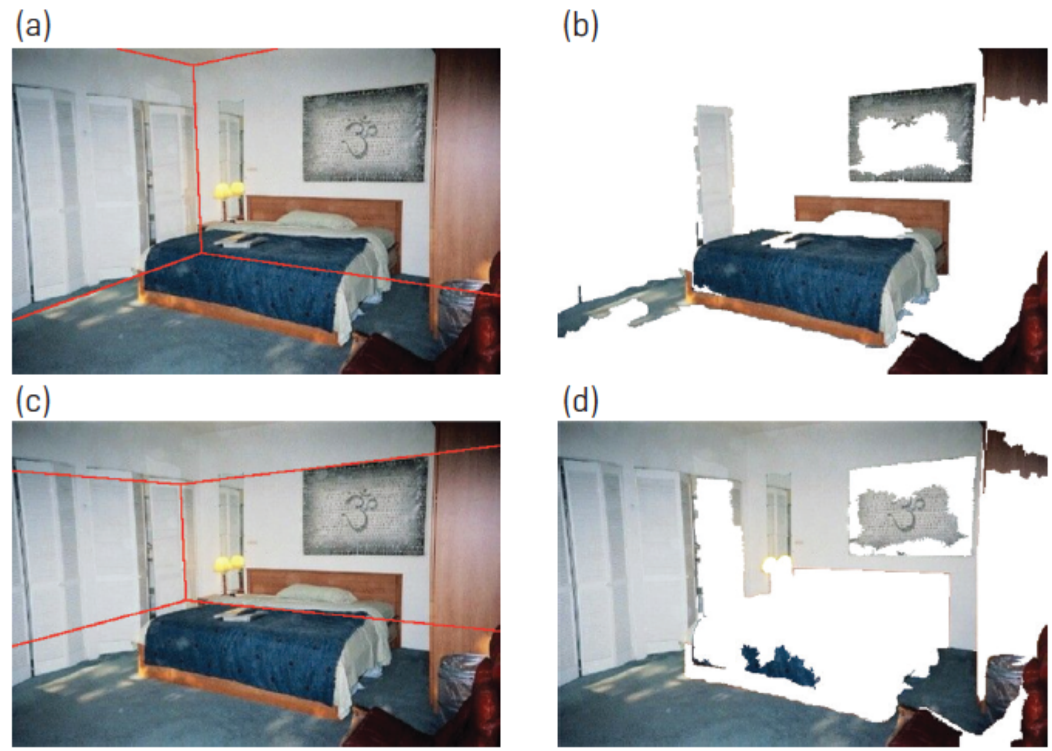
Figure 1. Output of our method. First row: the inferred box layout illustrated by red lines representing face boundaries. Second row: the inferred clutter layout.

Wang et al 13

Continuous y (?)

Prior knowledge helps, too

Figure 3. Example result of recovering box and clutter layout. The clutter layouts are shown by removing all non-clutter segments.
(a) Inferred box layout using model learned *with* prior knowledge.
(b) Inferred clutter layout using model learned *with* prior knowledge.
(c) Inferred box layout using model learned *without* prior knowledge.
(d) Inferred clutter layout using model learned *without* prior knowledge.



- Penalize:
  - variance in face appearance
  - too much clutter on face

Continuous y (?)

Wang et al 13

Latent clutter improves performance

	Hoiem et al.[10]	Hedau et al.[7] without	Hedau et al.[7] with	Ours without	without prior	$h = 0$	$h = GT$	cheat
Pixel	28.9%	26.5%	21.2%	*20.1 ± 0.5%*	21.5 ± 0.7%	22.2 ± 0.4%	24.9 ± 0.5%	19.2 ± 0.6%
≤20%	–	–	–	*62 ± 3*	58 ± 4	57 ± 3	46 ± 3	67 ± 3
≤10%	–	–	–	*30 ± 3*	24 ± 2	25 ± 3	20 ± 2	37 ± 4

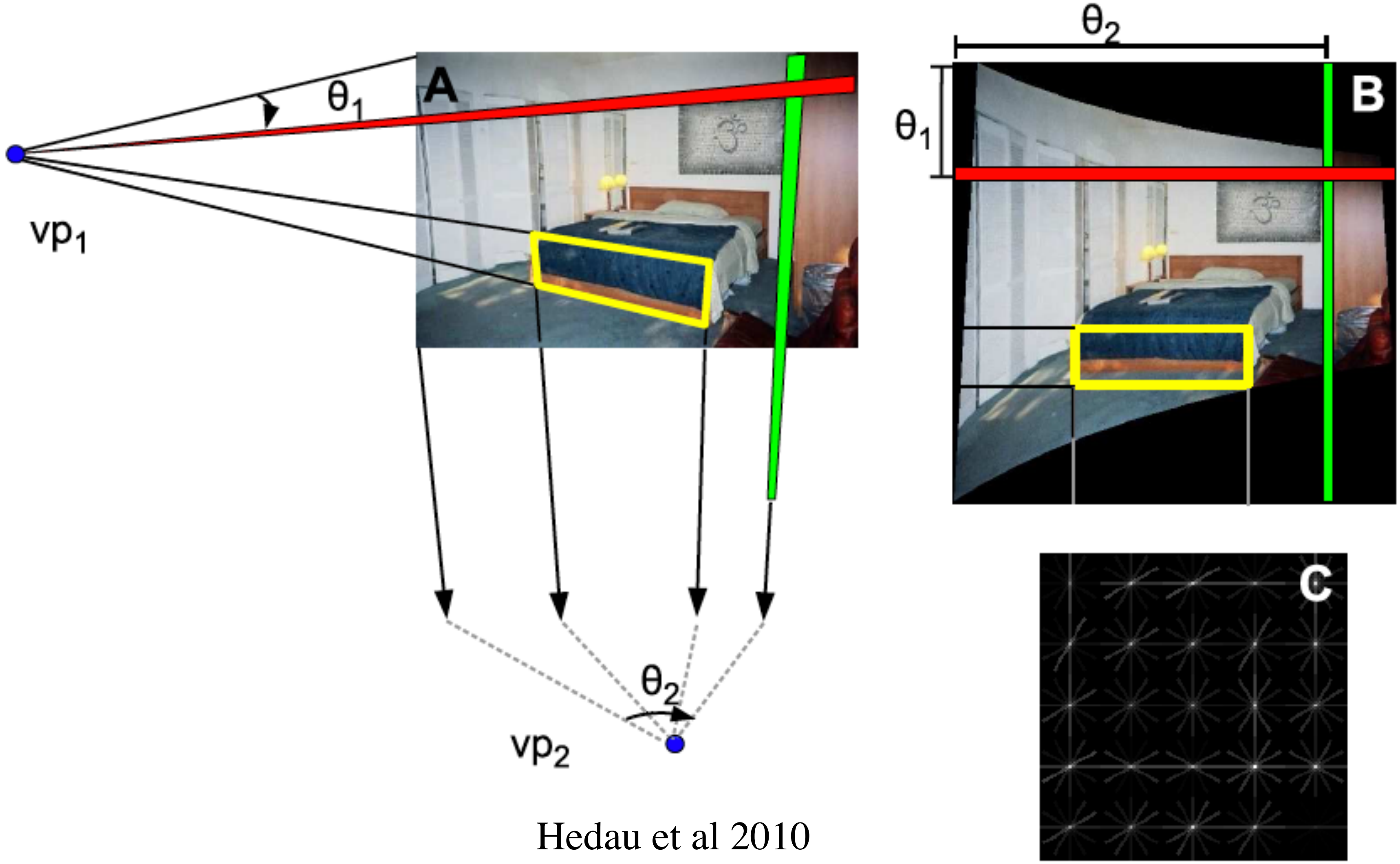
Accuracy

Wang et al 13

More examples

Wang et al 13

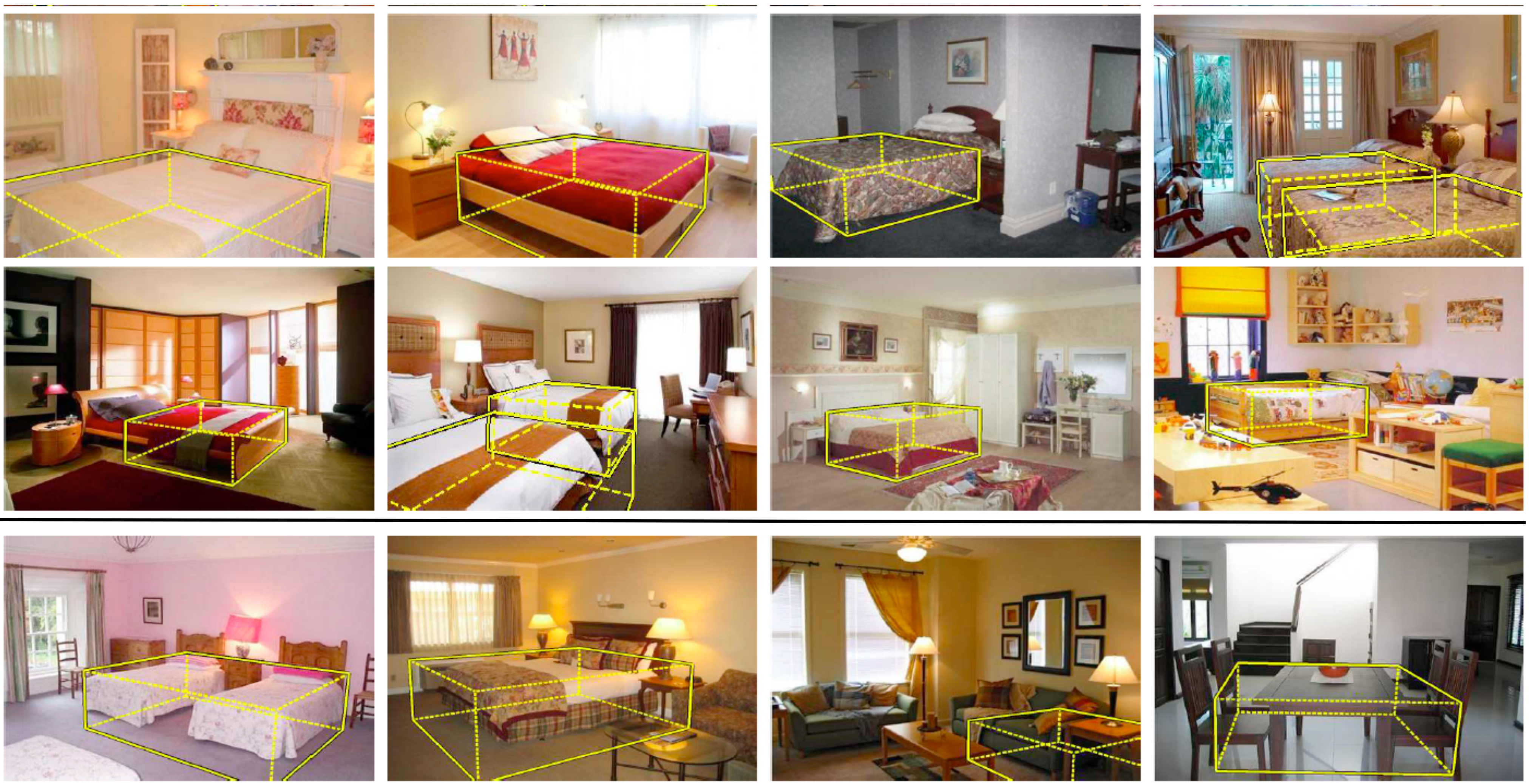
Detecting beds - I



Hedau et al 2010

Detecting beds - II

True positives



False positives

Hedau et al 2010

Detecting beds - III

- Beds constrain rooms
  - are axis-aligned
  - can’t pierce walls
- Variants
  - Box only (OK)
  - Box + 2D (better)
  - Jointly estimate room box, bed box(es) (best)

Hedau et al 2010

Joint estimation helps

Hedau et al 2010

Box-in-box gives accuracy improvements

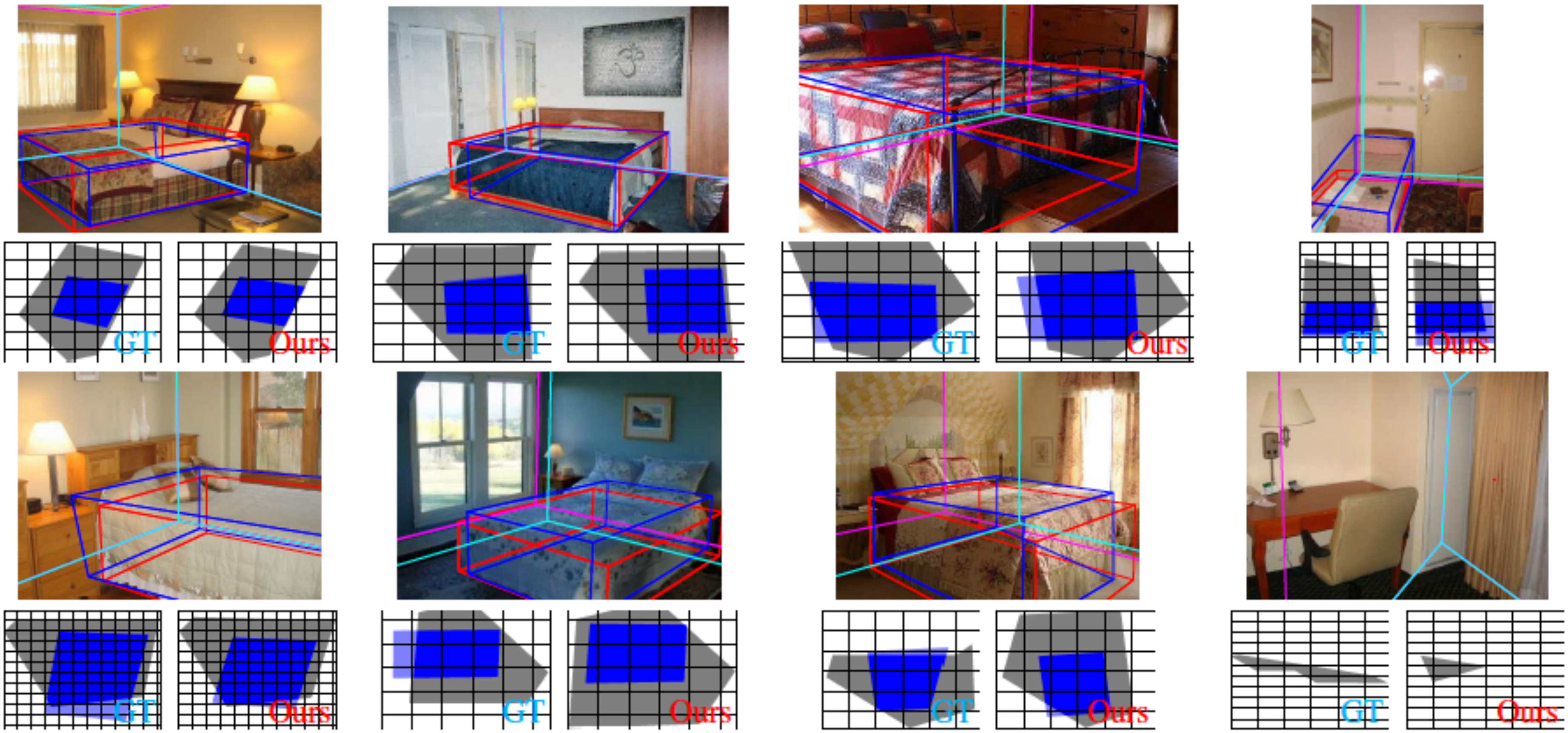


Figure 5. Illustration of prediction results (red, magenta) and best found ground truth state (blue, cyan) given vanishing points for joint object and layout inference overlaying the image. Below each image we provide visible annotation floor plan (gray) and object on the left while corresponding prediction result on the right. A failure case due to wrong vanishing points is illustrated in bottom right figure.

Schwing et al 13

Greedy application yields multiple boxes

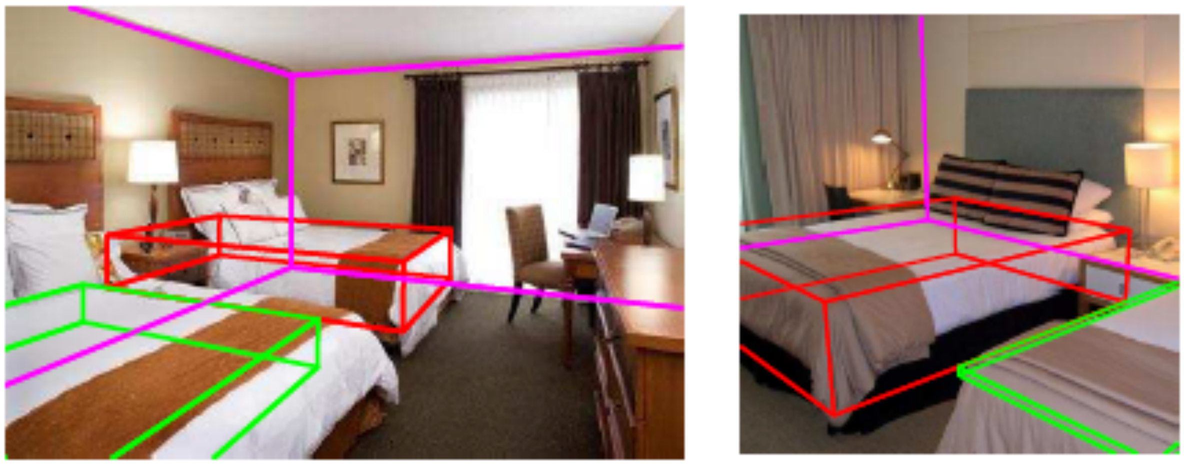


Figure 6. After jointly inferring layout (magenta) and object (red), we re-apply the object part to obtain a second object (green).

Schwing et al 13