

# Visual Understanding without Naming: Bypassing the “Language Bottleneck”



Alexei (Alyosha) Efros  
UC Berkeley

# Collaborators



Abhinav  
Gupta



Scott  
Satkin



David  
Fouhey



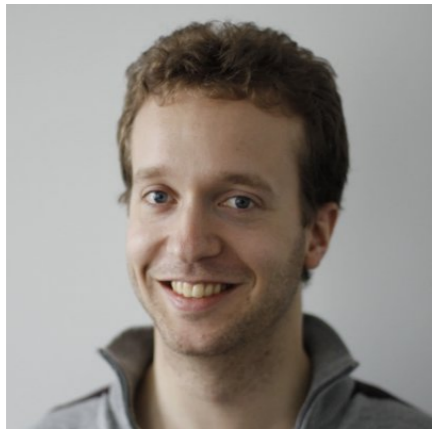
Martial  
Hebert



Natasha  
Kholgade



Yaser  
Sheikh



Vincent  
Delaitre



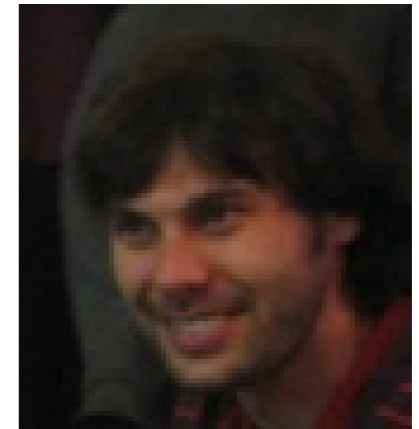
Mathieu  
Aubry



Bryan  
Russell



Ivan  
Laptev



Josef  
Sivic



Jun-Yan Zhu



Yong Jae Lee

# What do we mean by Visual Understanding?



slide by Fei Fei, Fergus & Torralba

# Object naming -> Object categorization



sky

building

flag

face

banner

wall

street lamp

bus

bus

cars

slide by Fei Fei, Fergus & Torralba

# Image Labeling

sky

building

flag

face

banner

wall

street lamp

bus

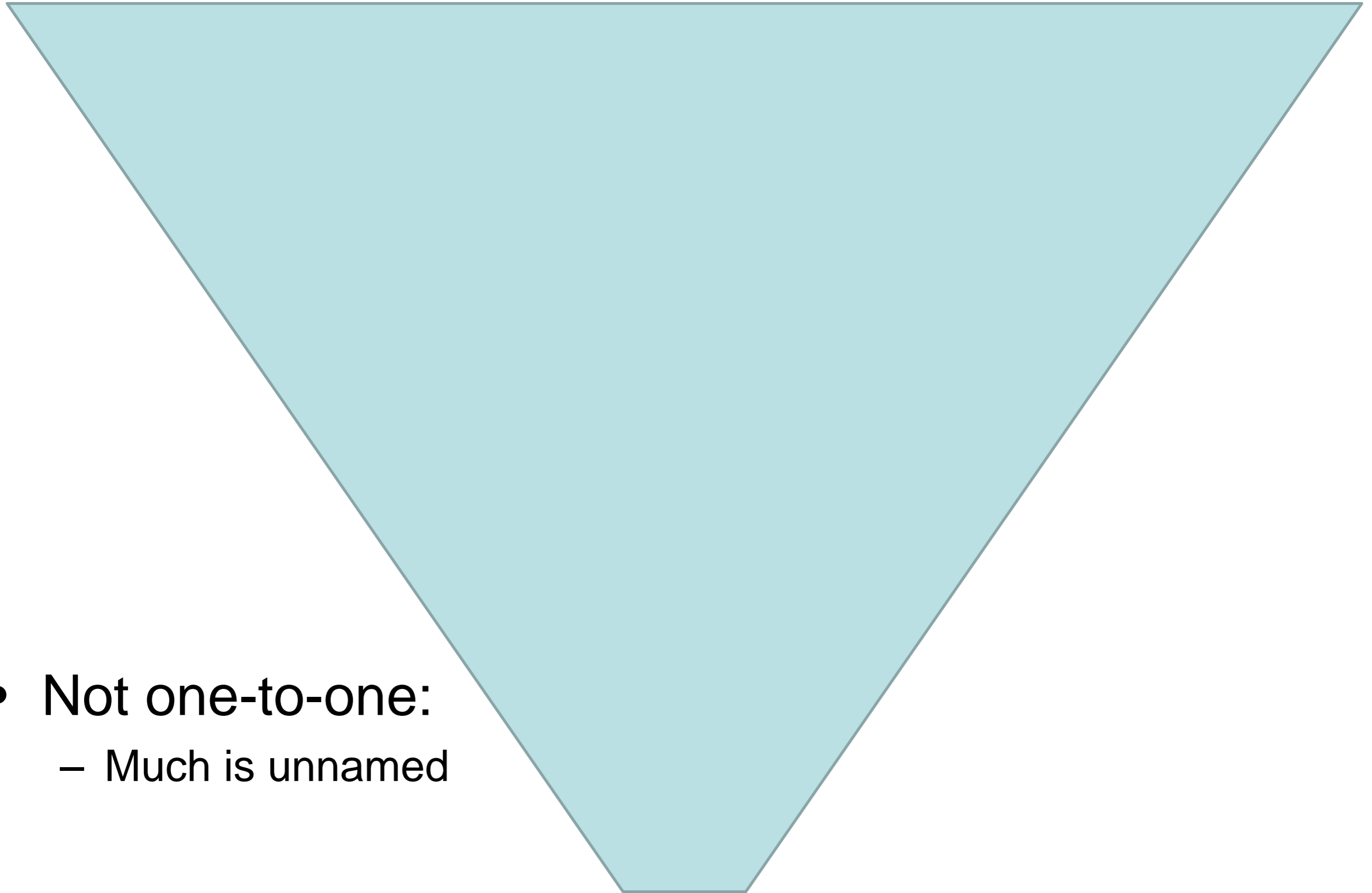
bus

cars



Hays and Efros, "Where in the World?", 2009

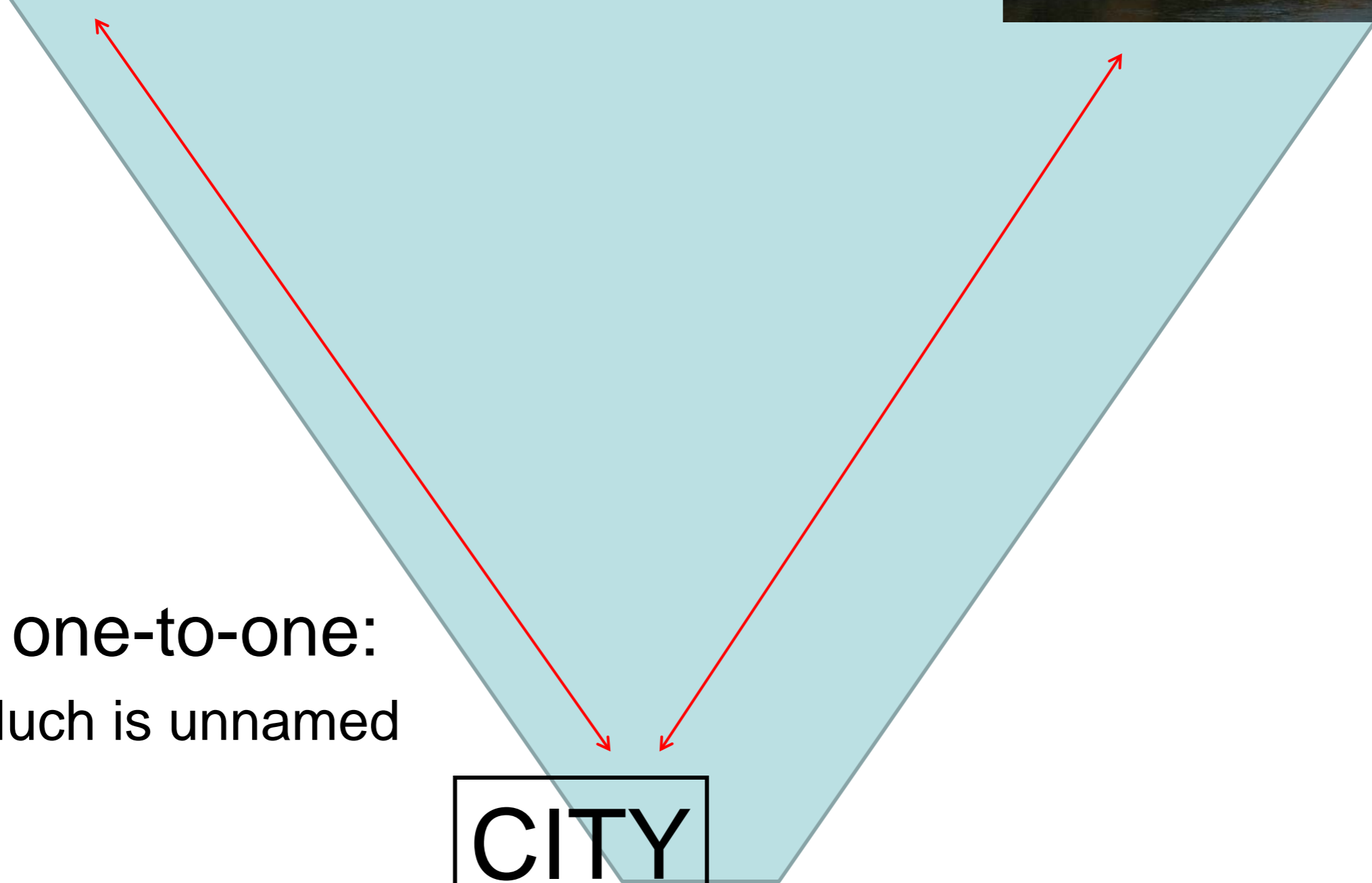
# Visual World



words

- **Not one-to-one:**
  - Much is unnamed

# Visual World



**CITY**

words

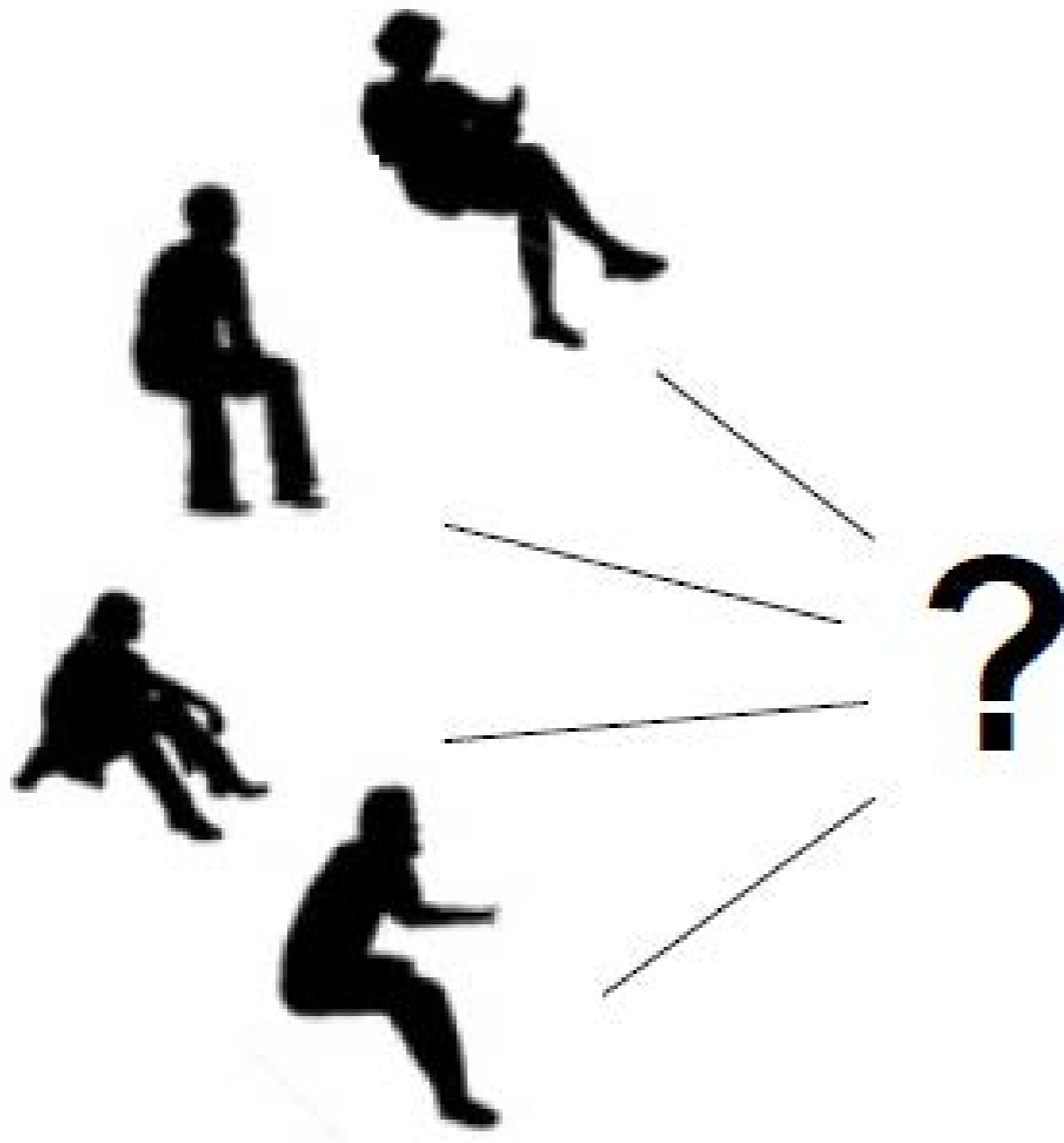
- **Not one-to-one:**
  - Much is unnamed



# Verbs (actions)

**sitting**

# Visual “sitting”

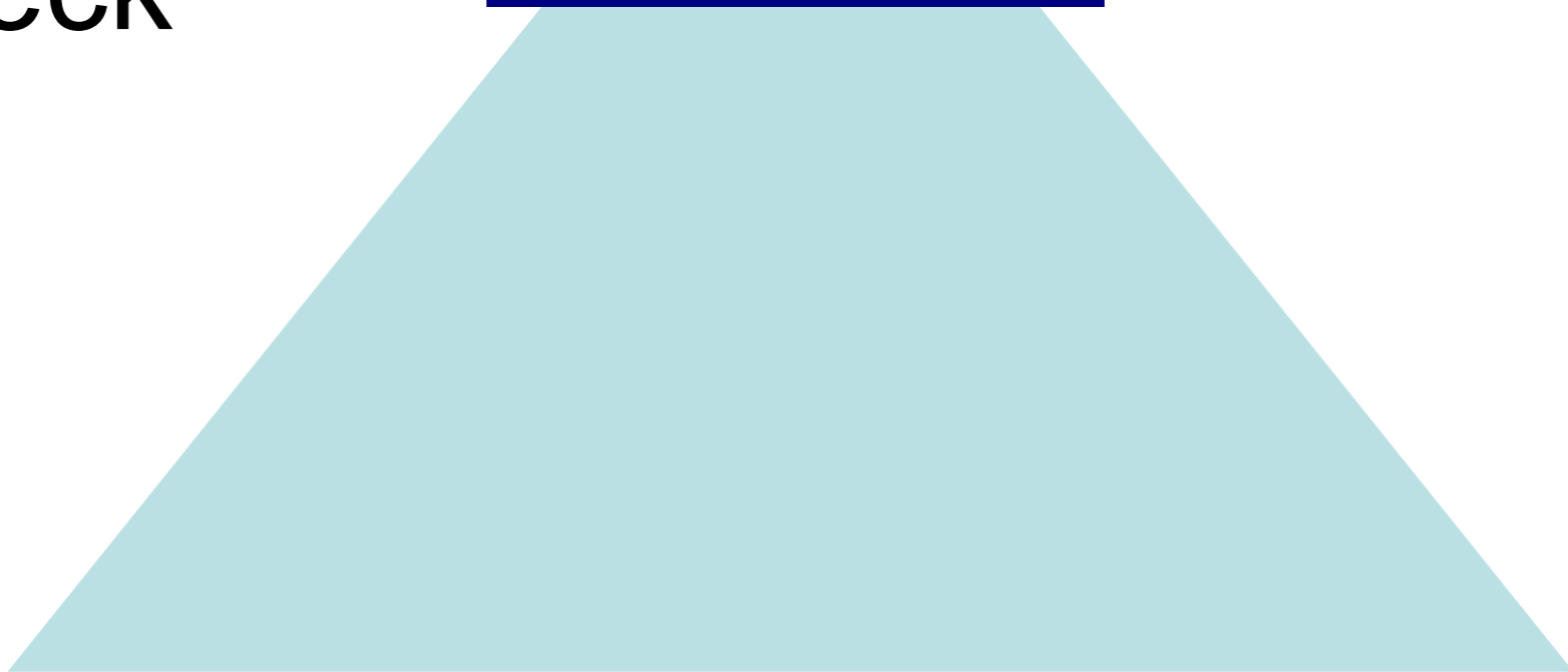


Visual World



**words**

The Language  
Bottleneck



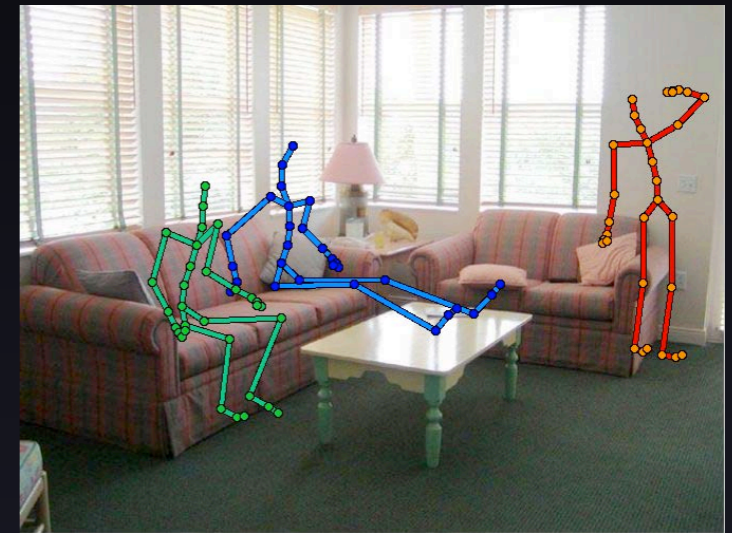
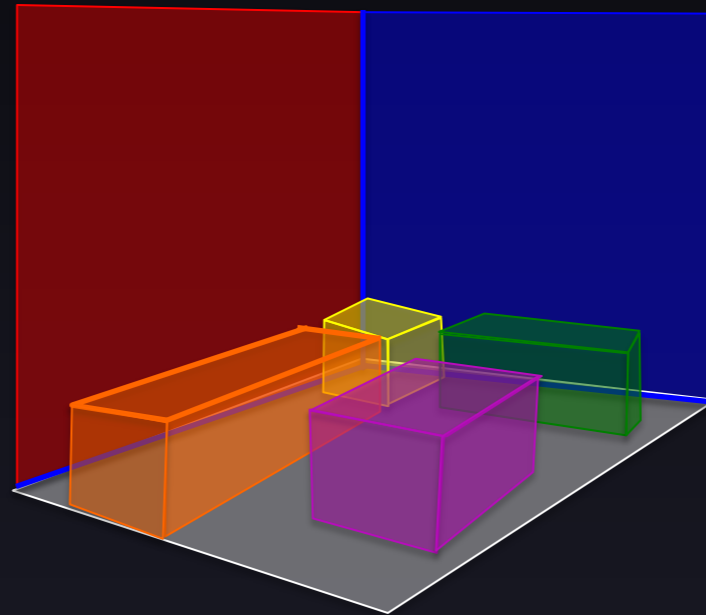
Scene understanding, spatial reasoning, prediction,  
image retrieval, image synthesis, etc.

# Visual World

- 1. 3D Human Affordances**
- 2. 3D Object Correspondence**
- 3. User-in-the-visual-loop**

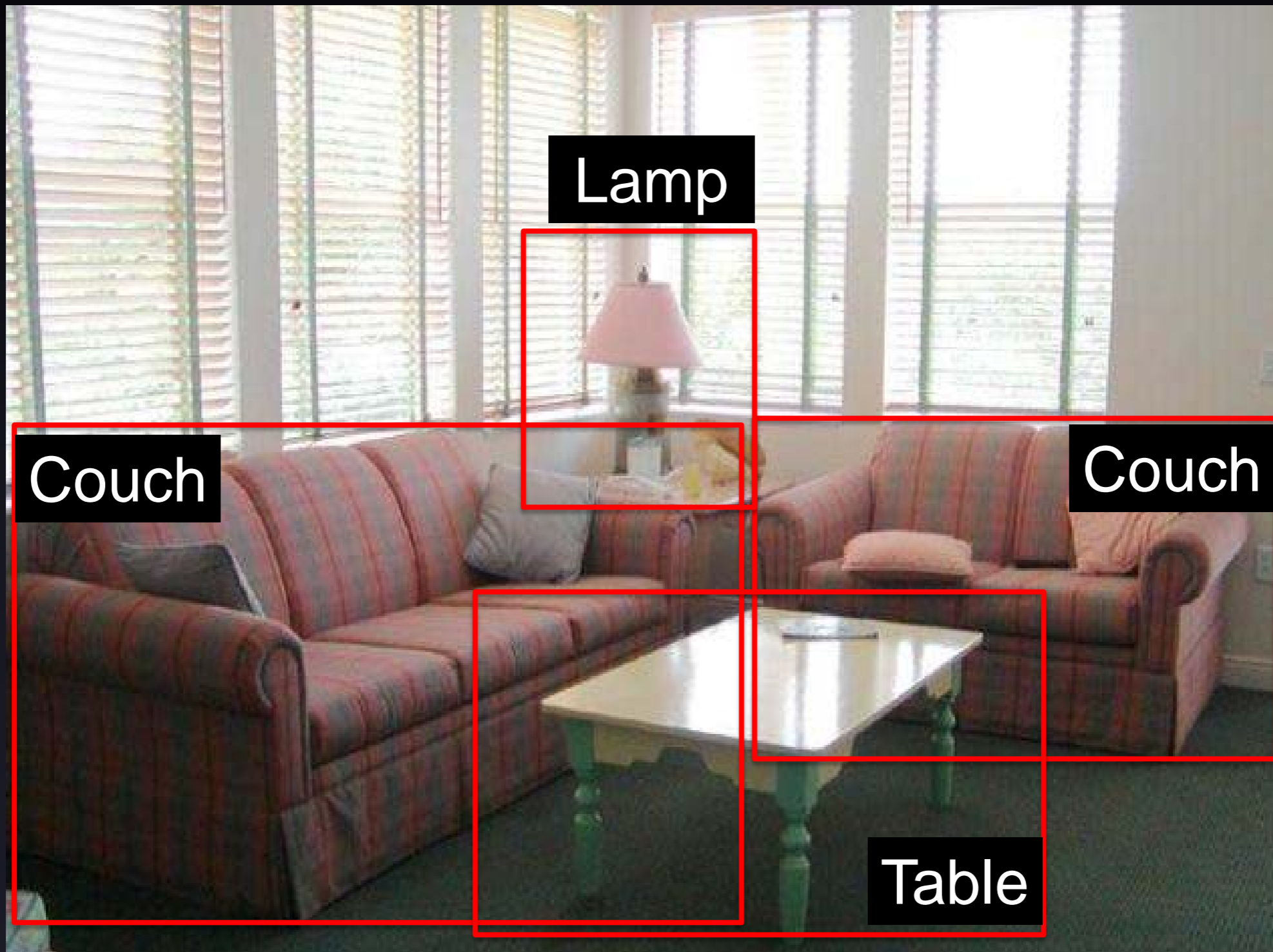
Scene understanding, spatial reasoning, prediction, image retrieval, image synthesis, etc.

# From 3D Scene Geometry to Human Workspaces

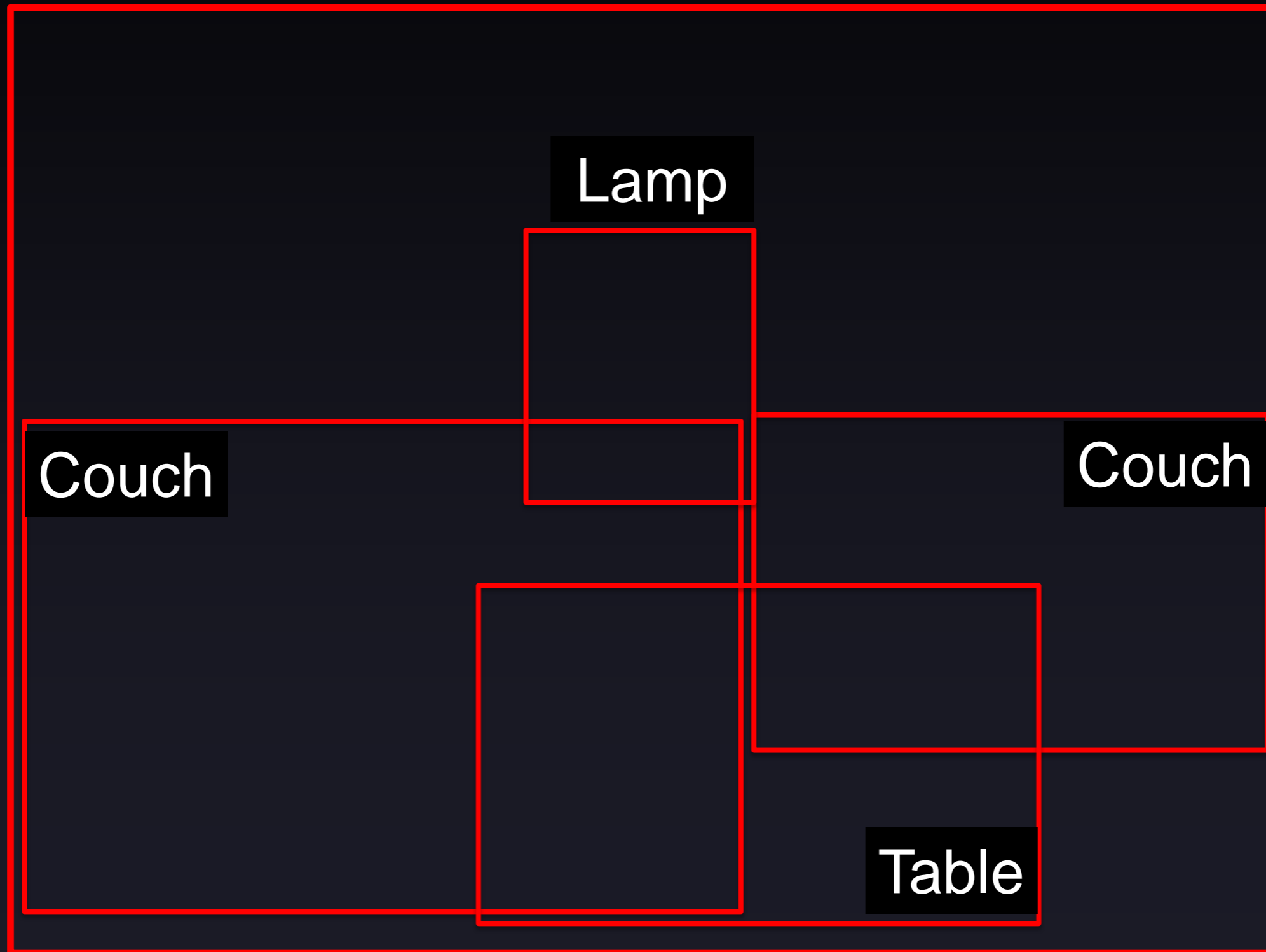


Abhinav Gupta, Scott Satkin, Alexei Efros and Martial Hebert  
CVPR'11

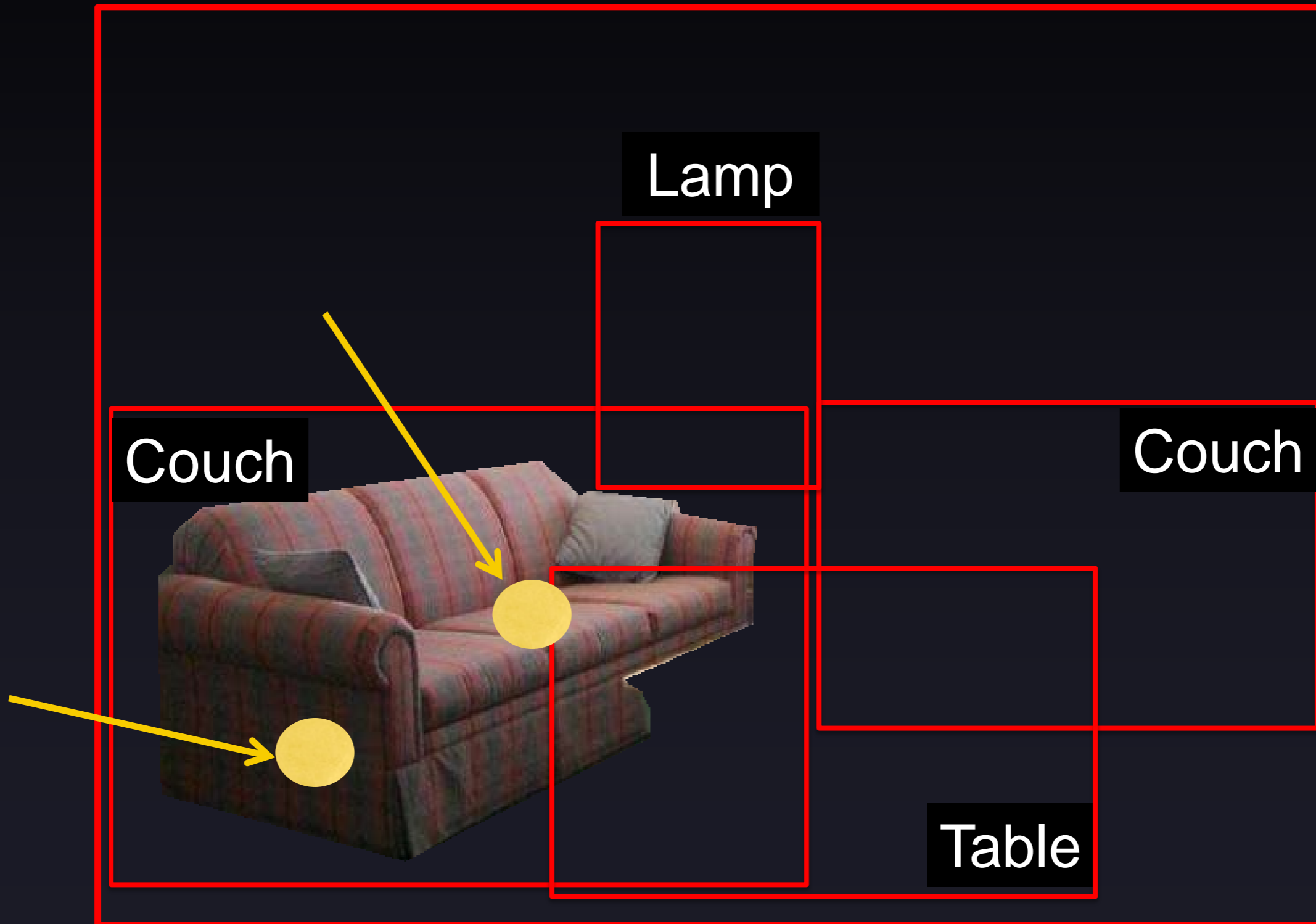
# Object Naming



Is there a **couch** in the image?

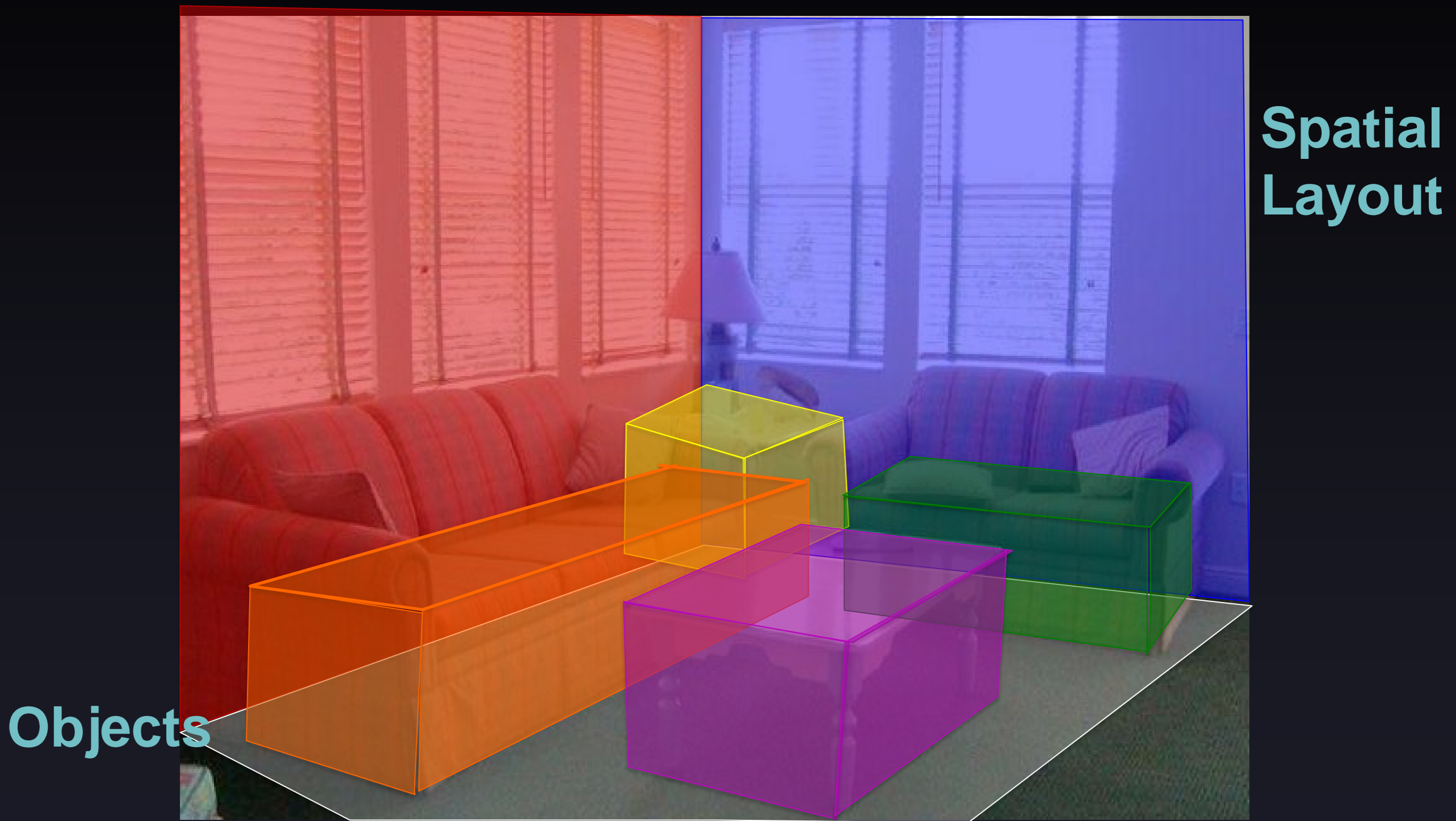


# Where can I **sit** ?



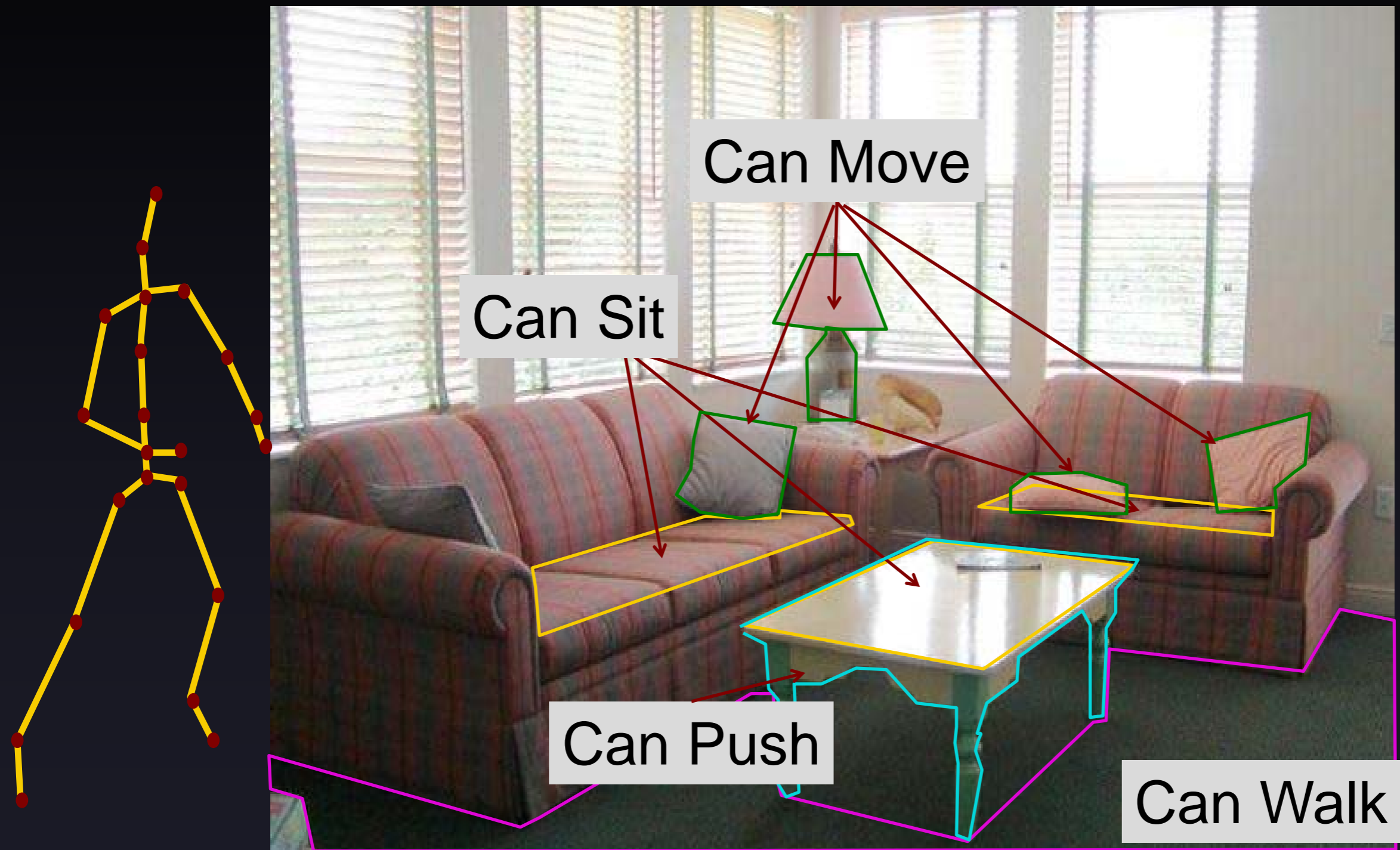


# 3D Indoor Image Understanding



Hoiem et al. IJCV'07, Delage et al. CVPR'06, Hedau et al. ICCV'09., Lee et al. NIPS'10, Wang et al. ECCV'10

# Human Centric Scene Understanding



Reasoning in terms of set of allowable actions



# Sitting



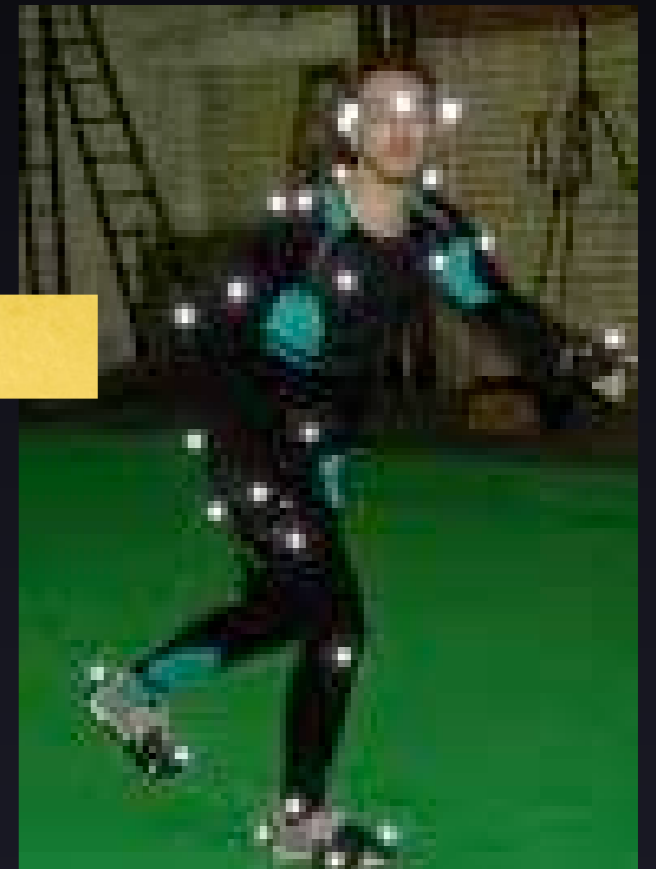
# Pose-defined Vocabulary



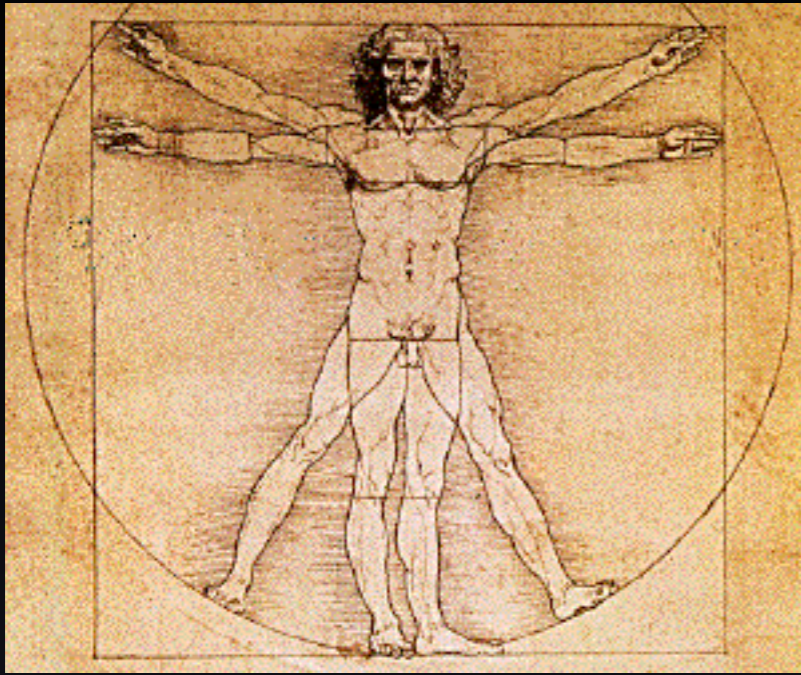
Sitting



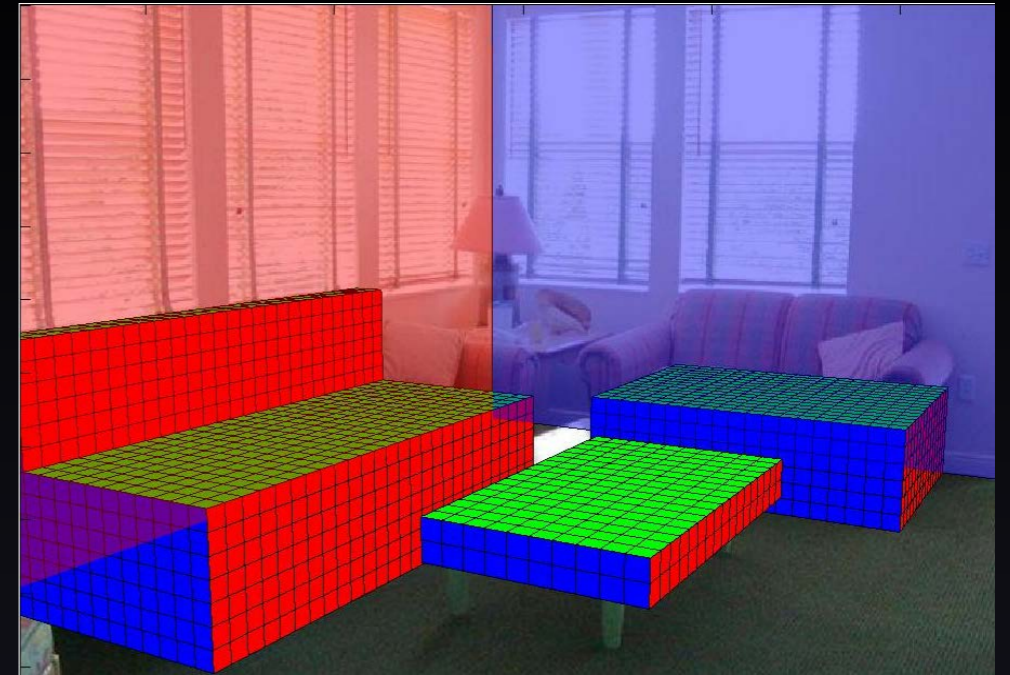
Poses



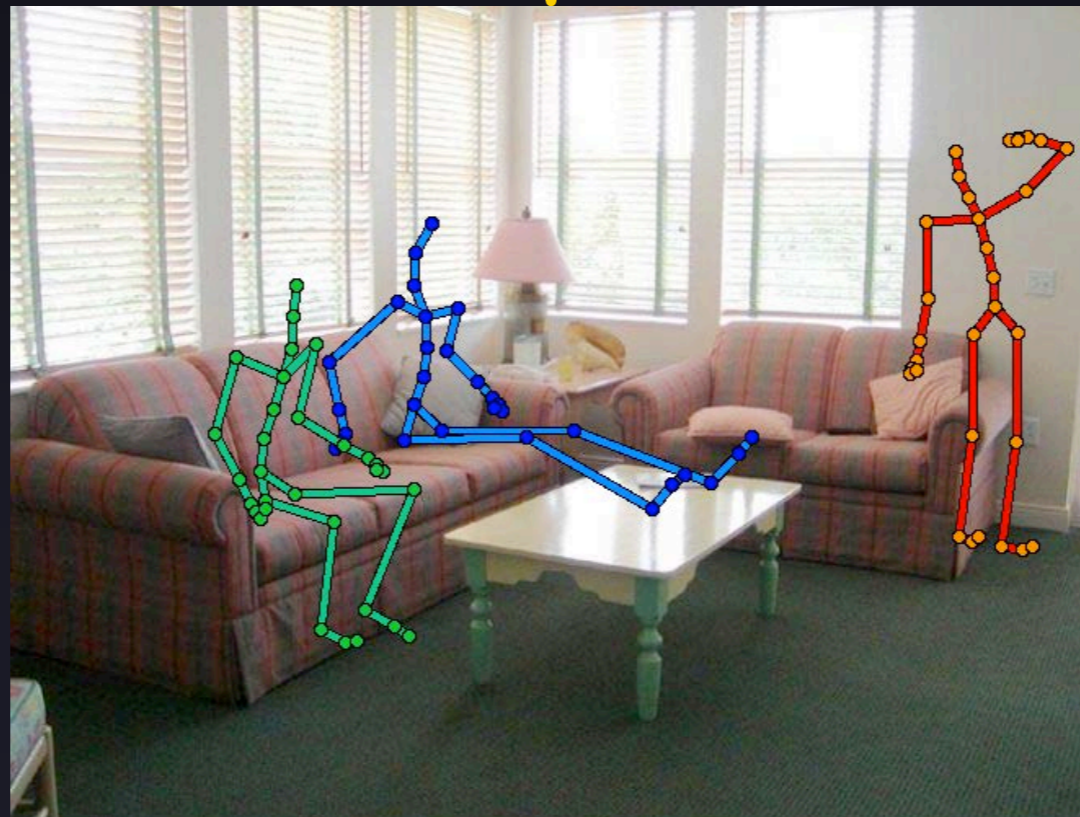
Motion  
Capture



Human Workspace

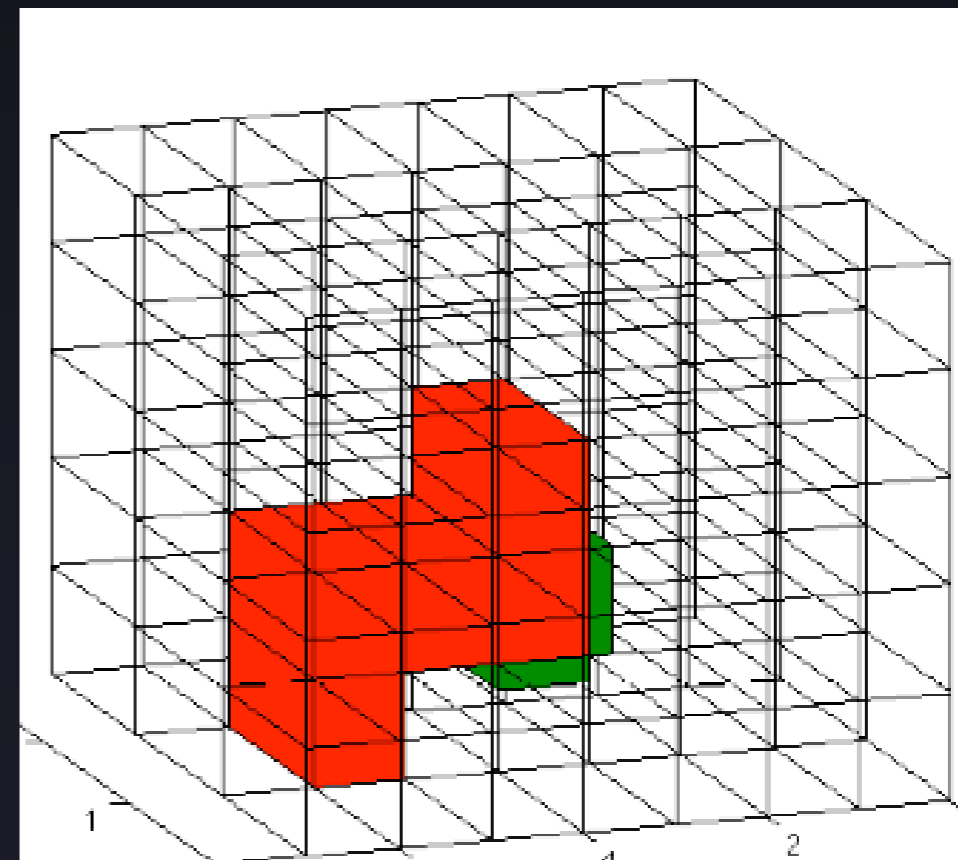
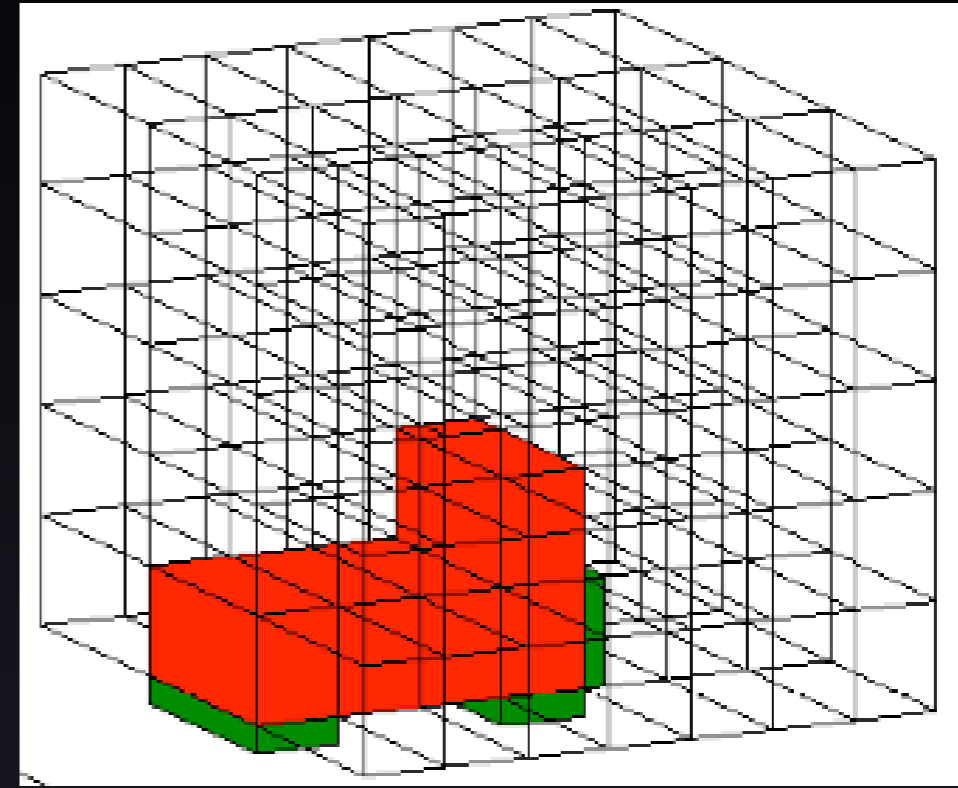
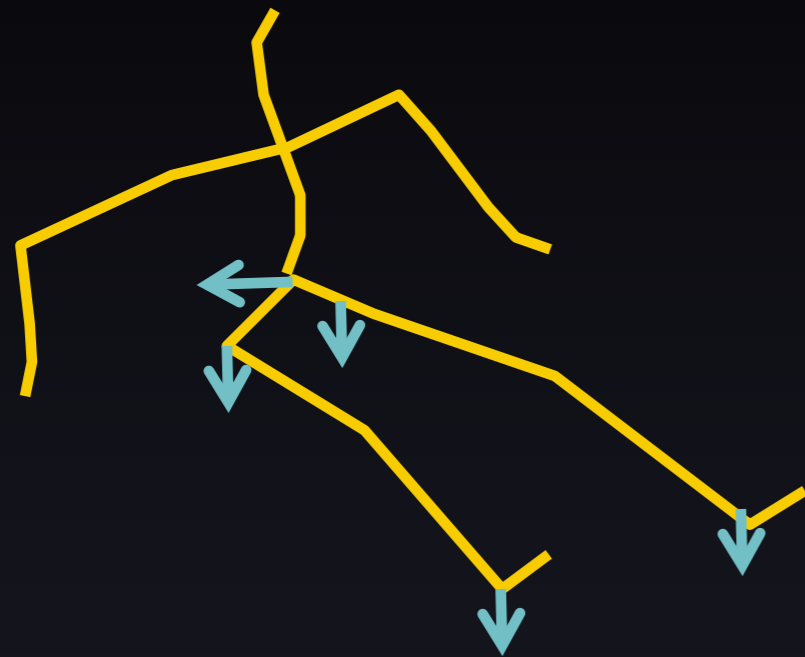


3D Scene Geometry



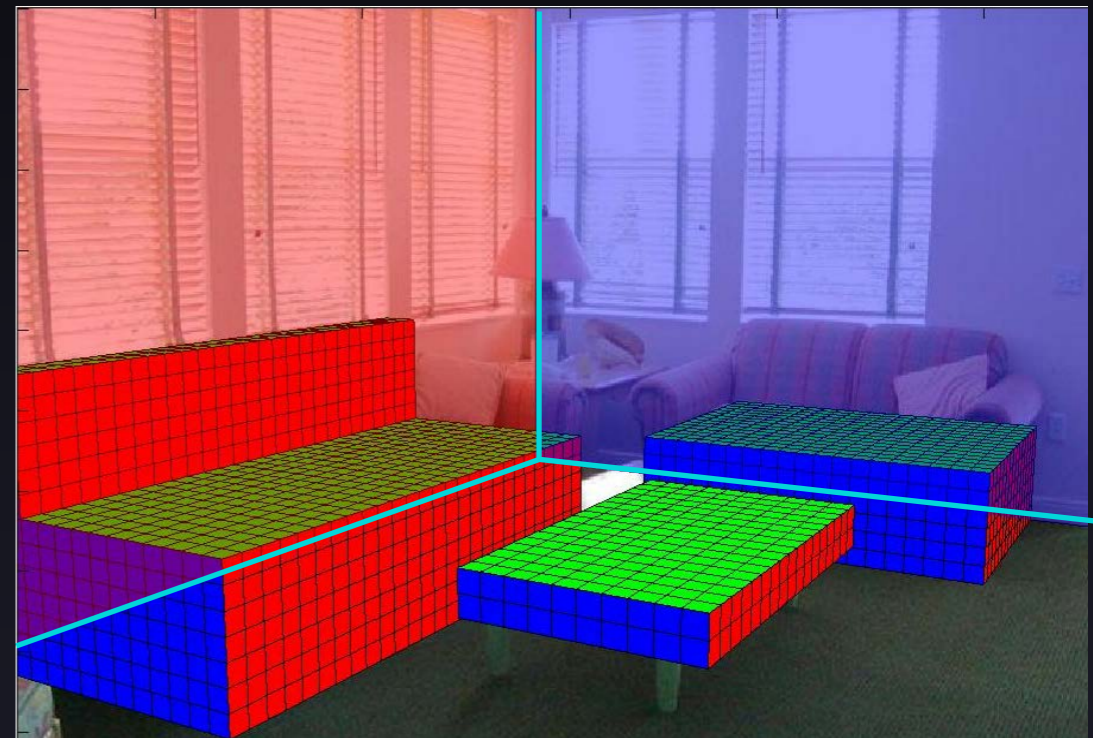
Joint Space of Human-Scene Interactions

# Qualitative Representation



# 3D Scene Geometry

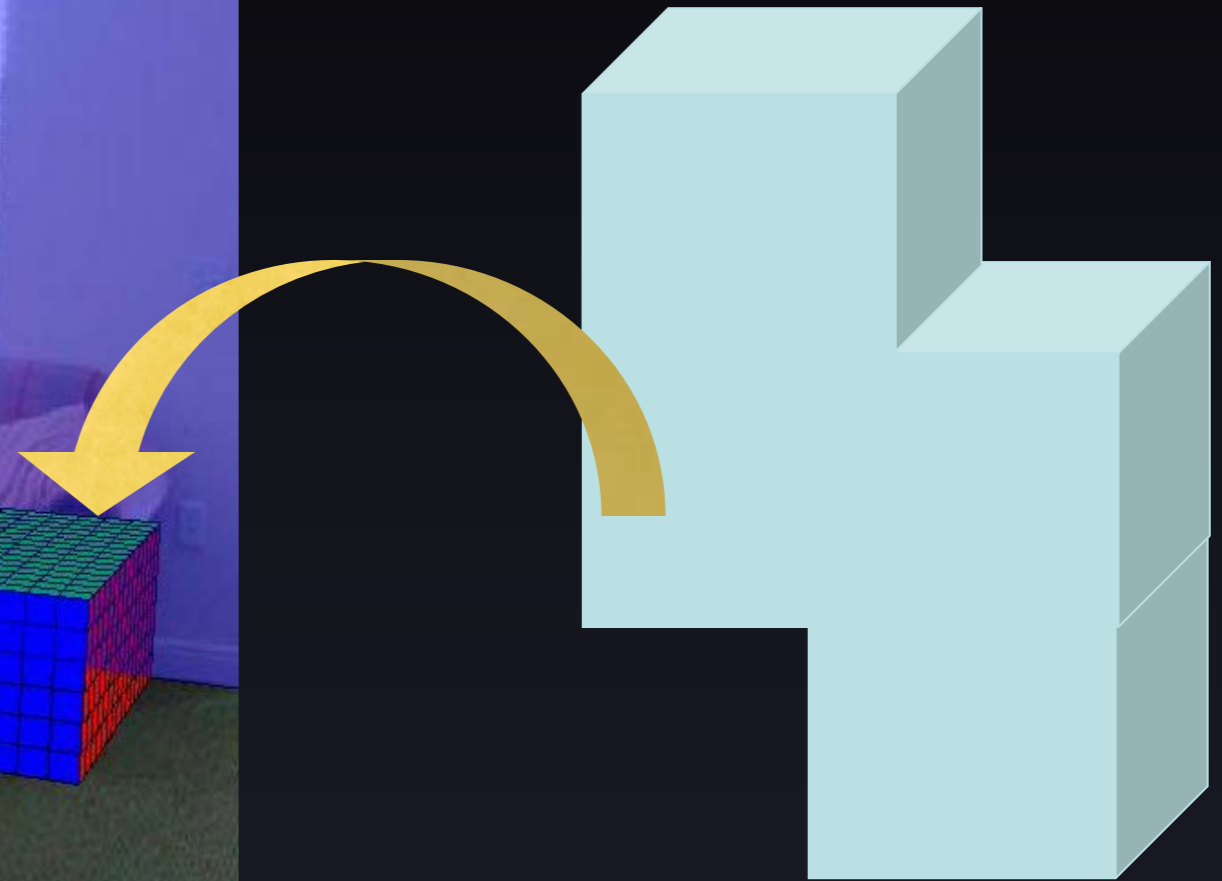
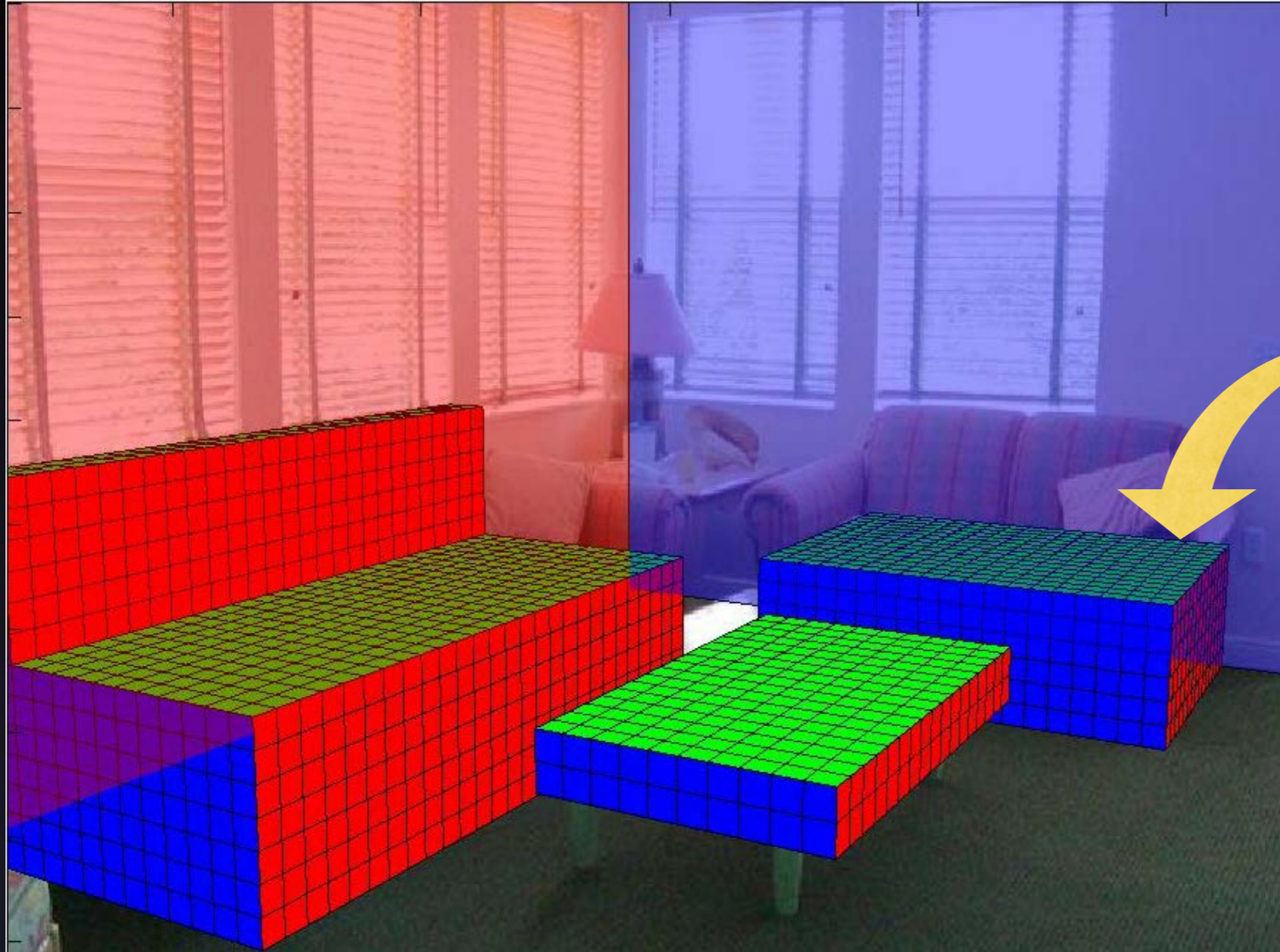
- Each scene modeled by
  - Layout of the Room
  - Layout of the Objects
- Room Represented by inside-out box
- Objects represented by occupied voxels.



## References:

Hedau et al. ICCV'09., Lee et al. NIPS'10, Wang et al. ECCV'10

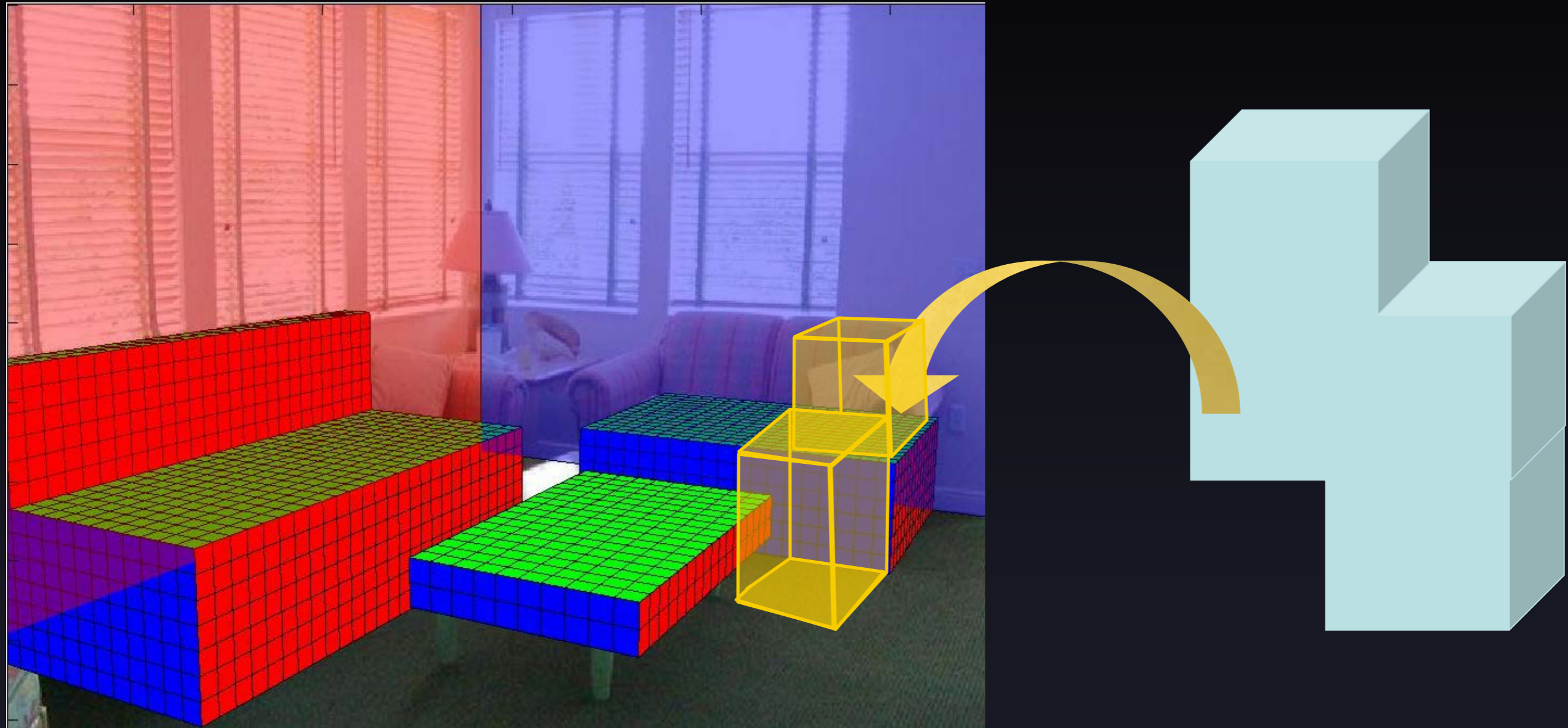
# Goal



Where would the Human Block fit ?

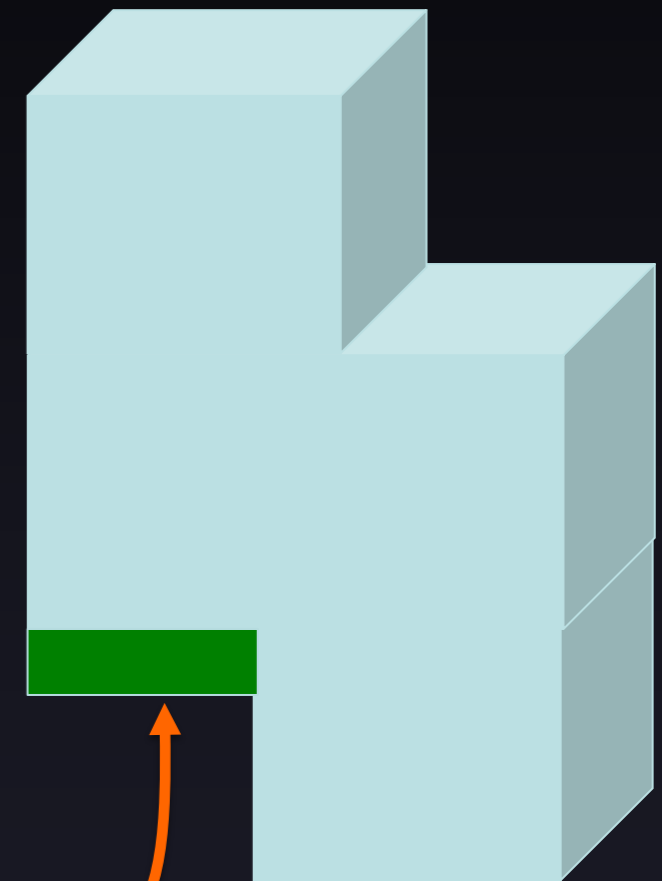
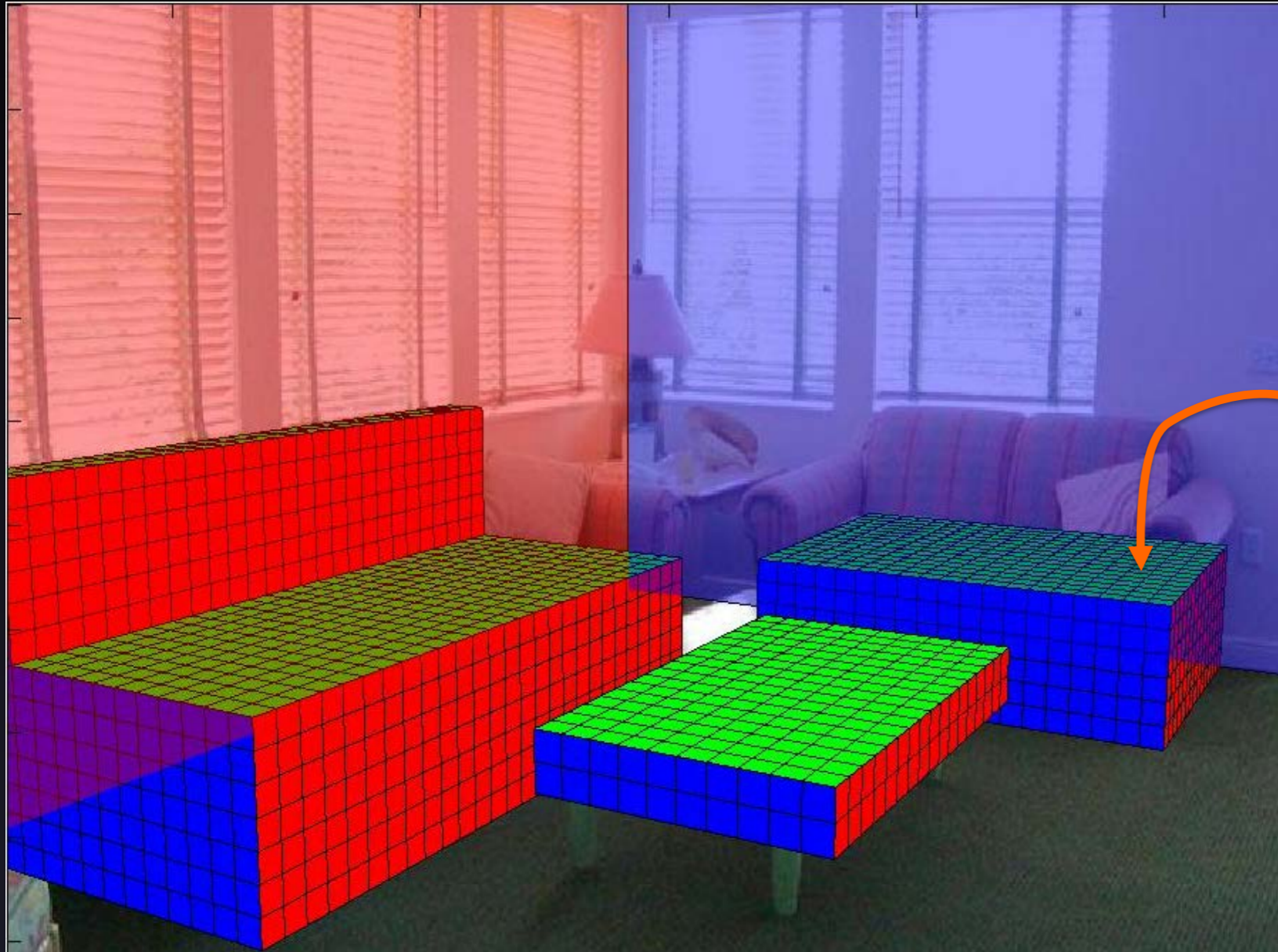


# Human Scene Interactions



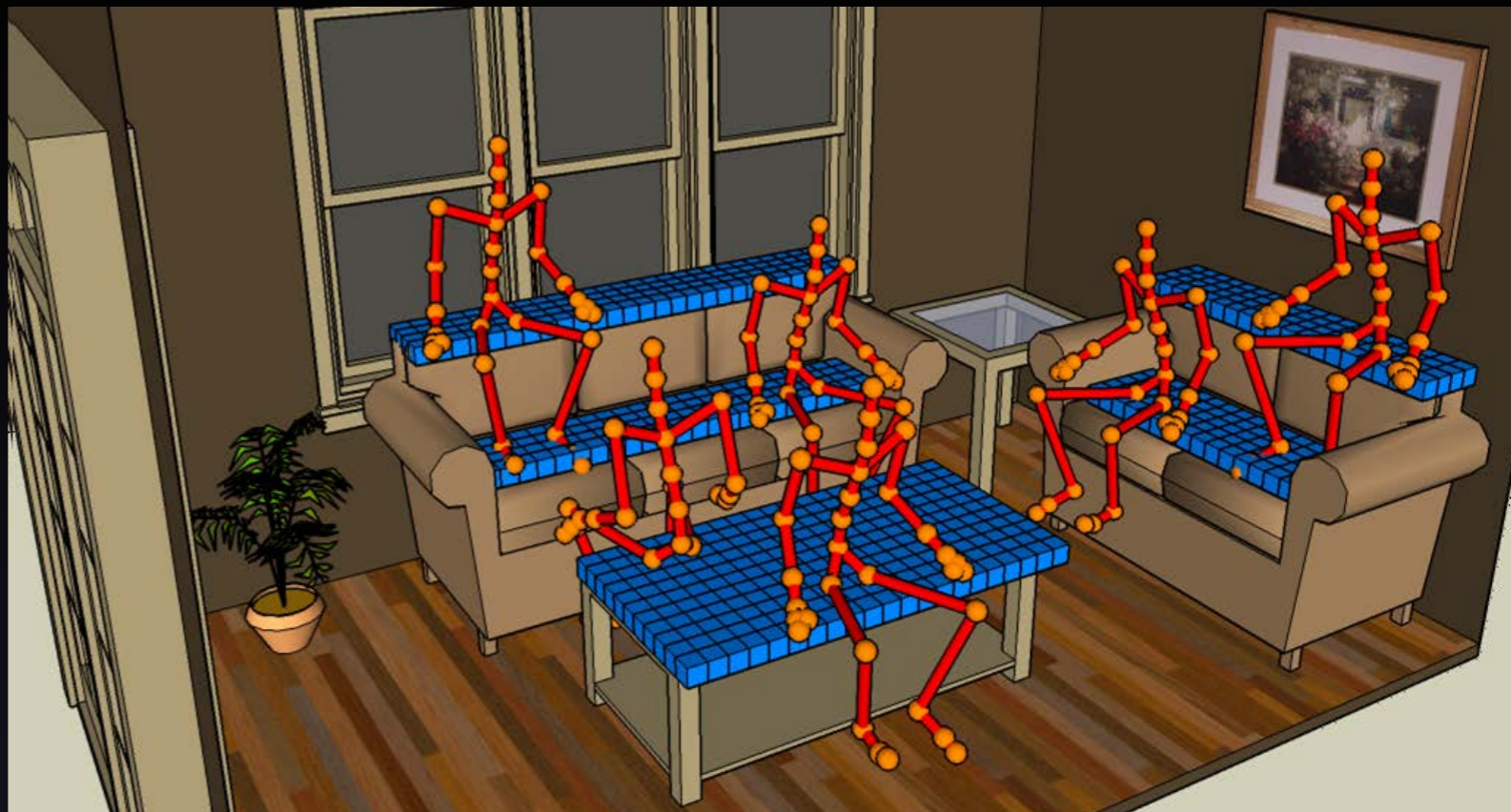
**Free Space Constraint** : No Intersection between Human Block and Objects

# Human Scene Interactions



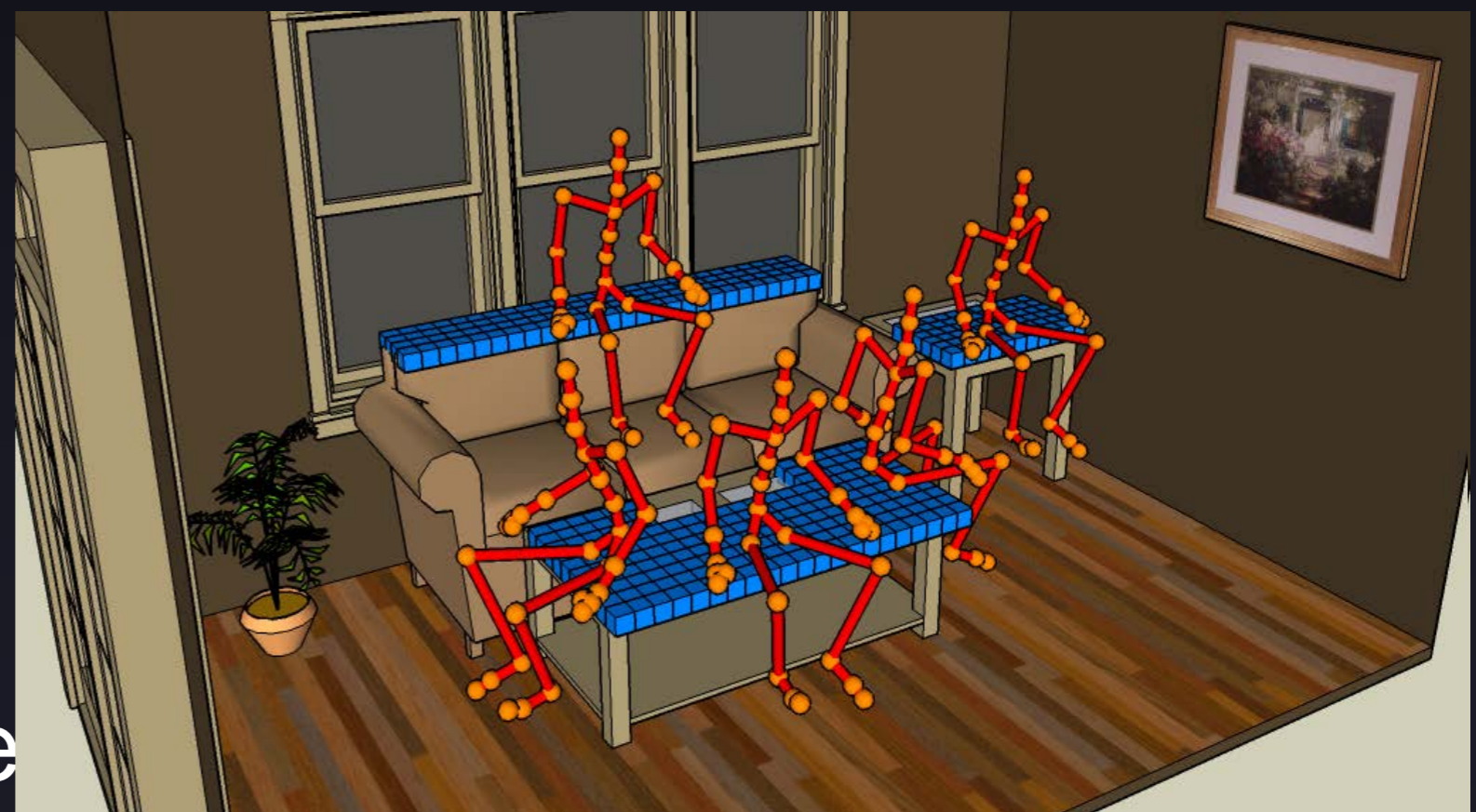
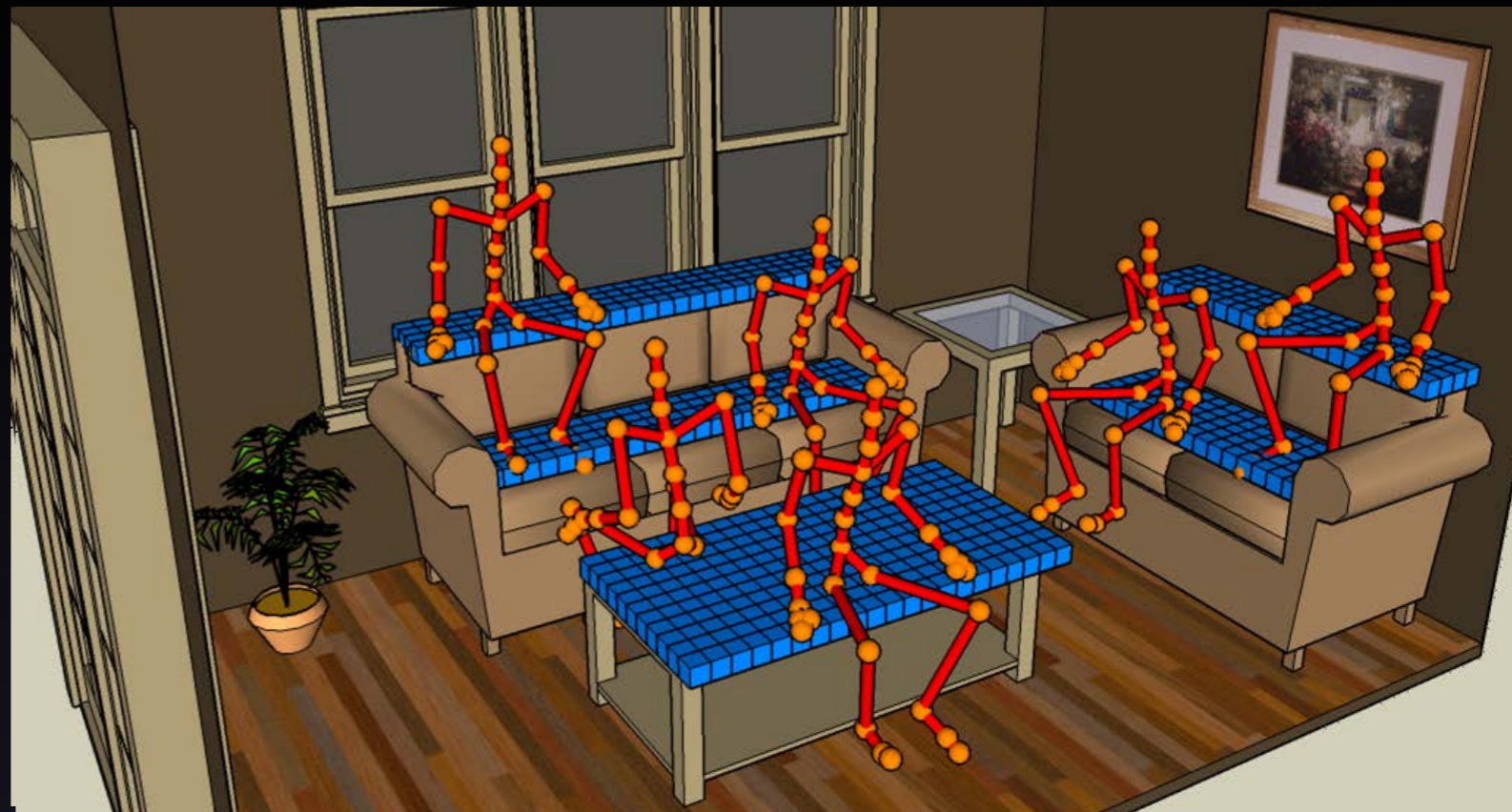
**Support Constraint : Presence of Objects for Interaction**

# Ground-Truth 3D Geometry



Data Source:  
Google 3D Warehouse

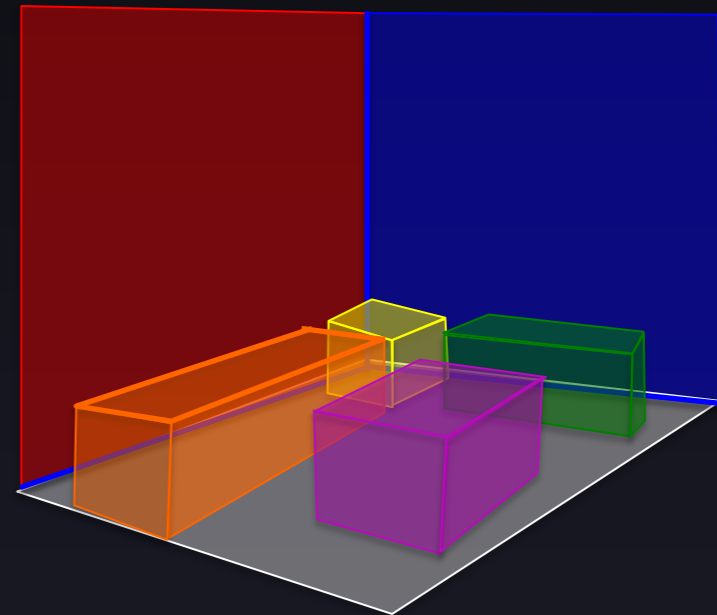
# Ground-Truth 3D Geometry



Data Source:  
Google 3D Warehouse

# Extracting 3D Geometry

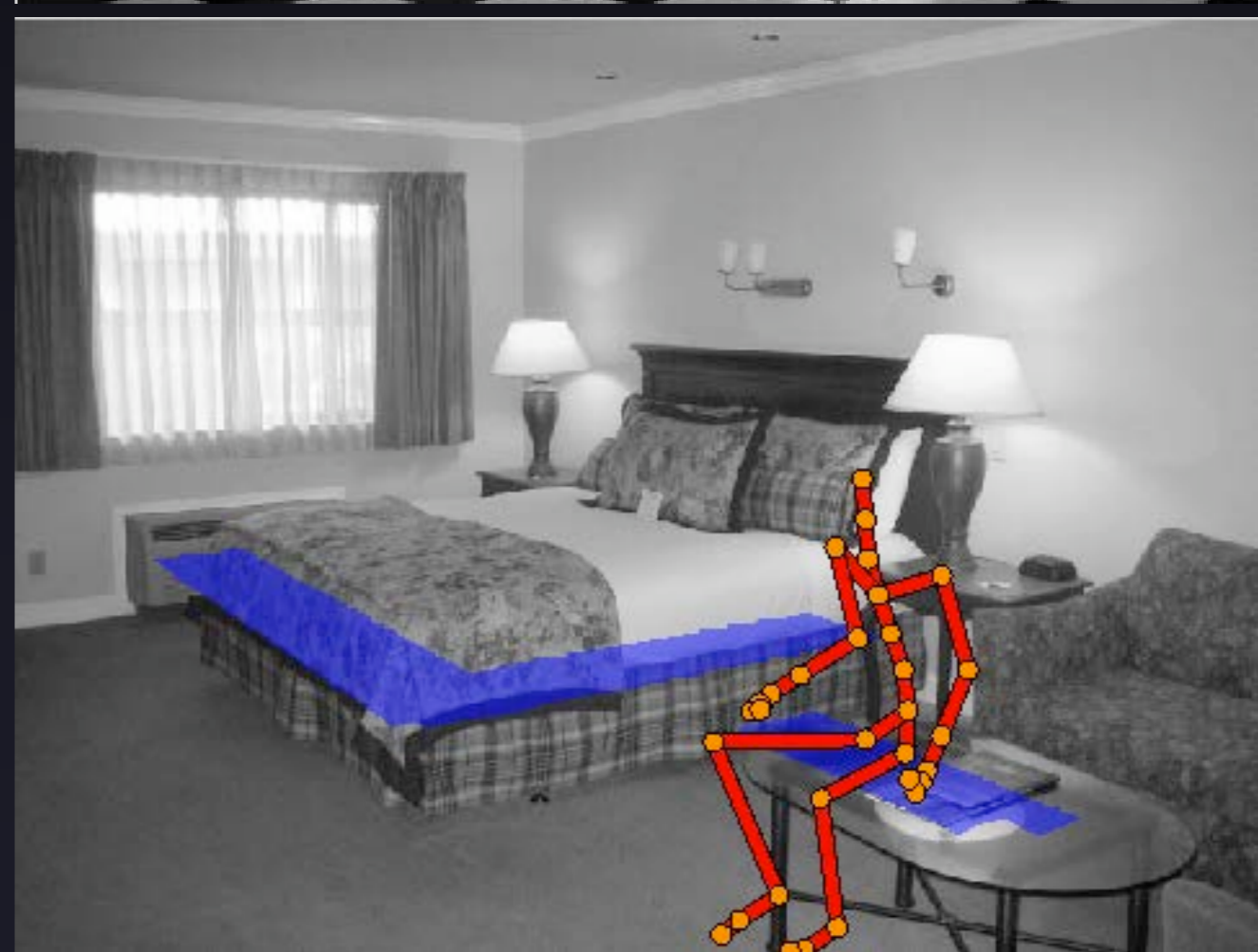
- Estimating 3D Scene Geometry from a single image is an extremely difficult problem.



- Build on work in 3D Scene Understanding of [Hedau'09] and [Lee'10]





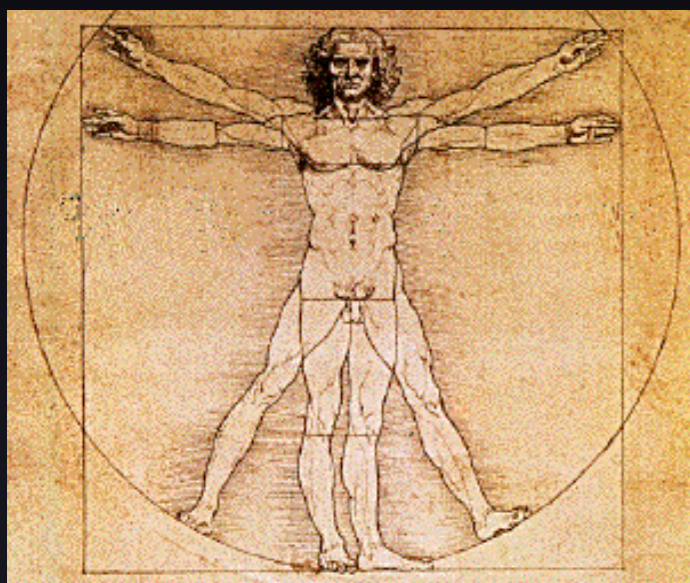




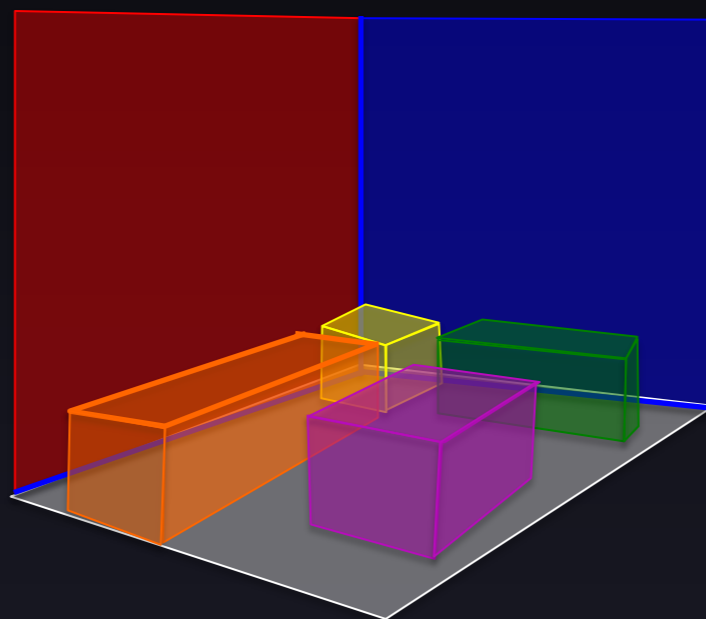
# Subjective Scene Interpretation



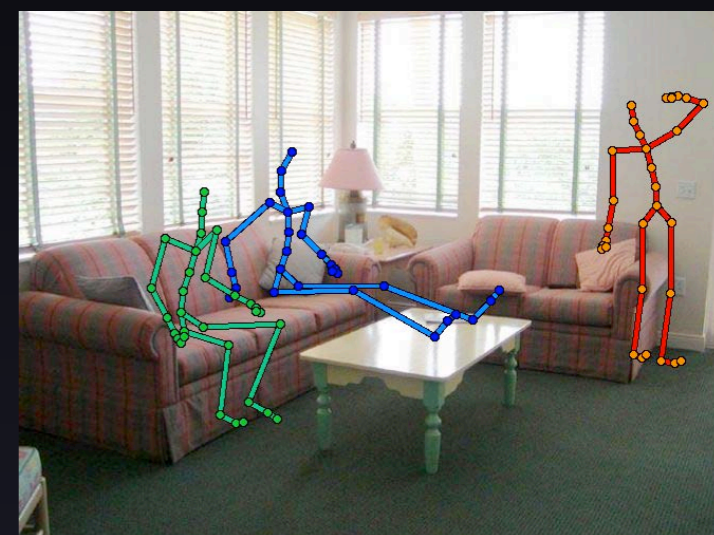
# Summary



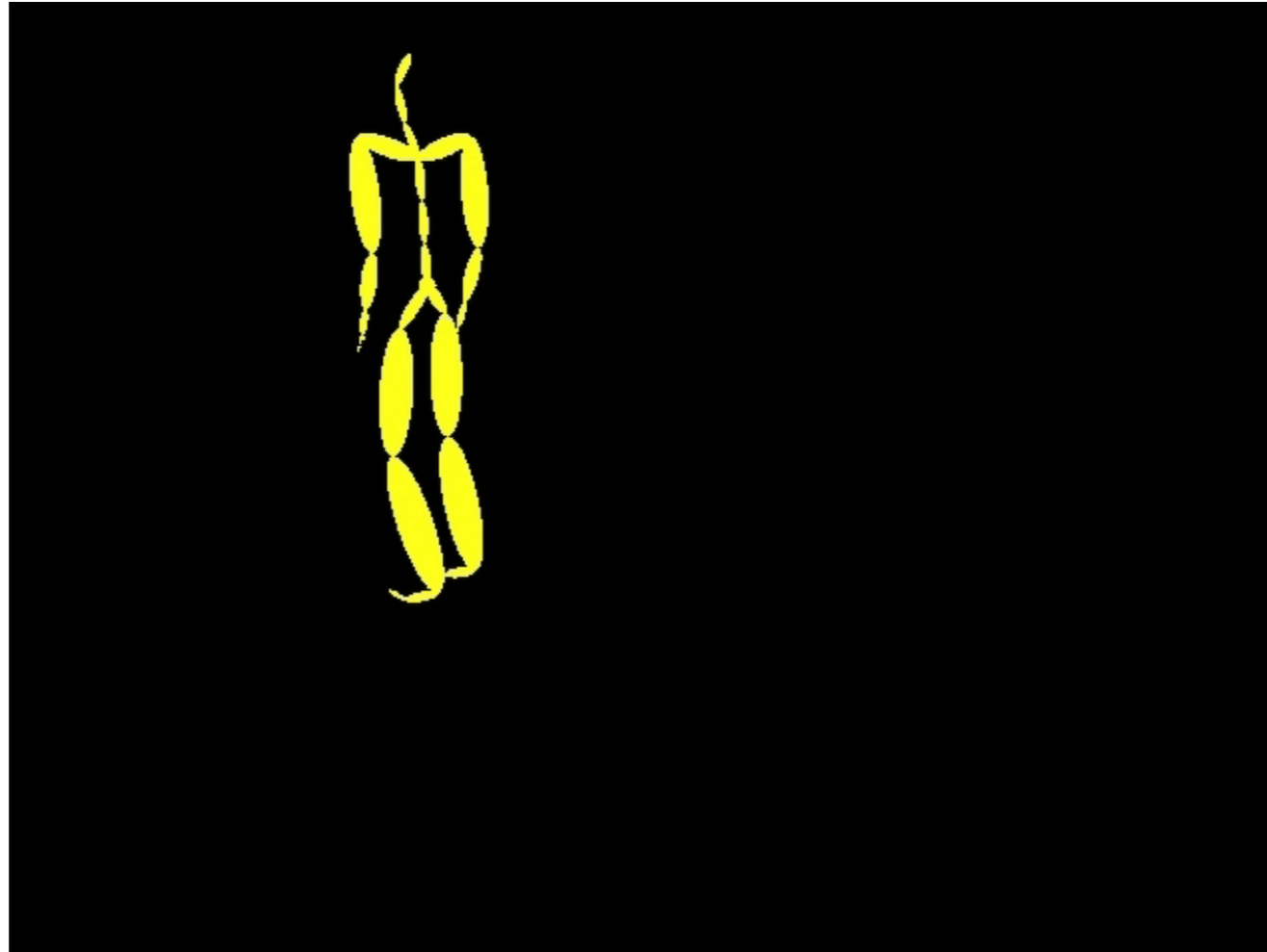
+



=



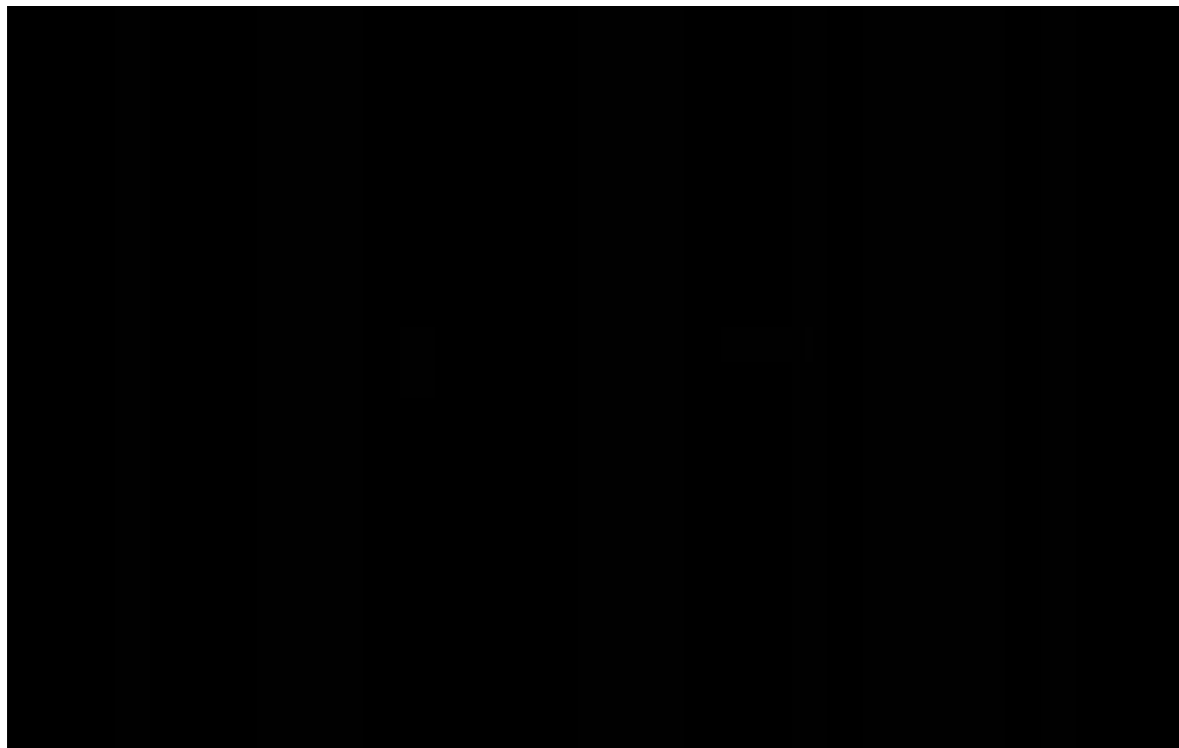
# The Inverse Problem



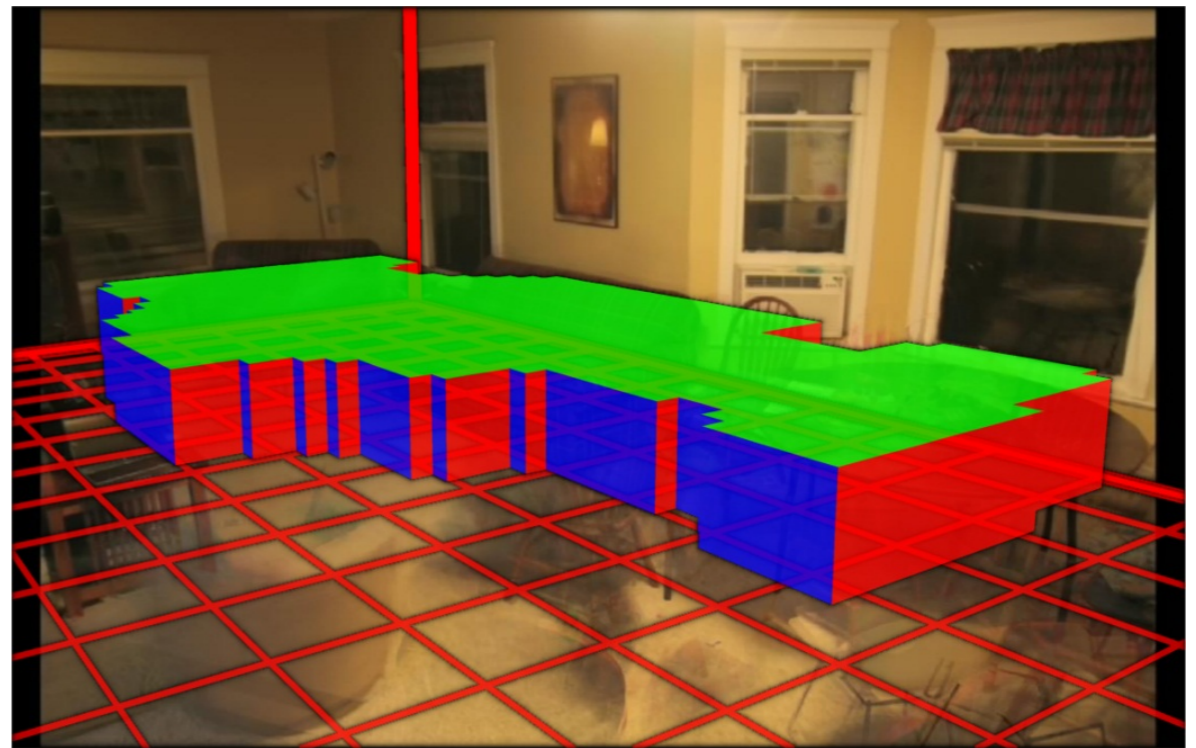
# People Watching: Human Actions as a Cue for Single-View Geometry

David Fouhey, Vincent Delaitre,  
Abhinav Gupta, Alexei Efros, Ivan Laptev, Josef Sivic  
ECCV 2012

# Humans as Active Sensors



Input:  
Timelapse

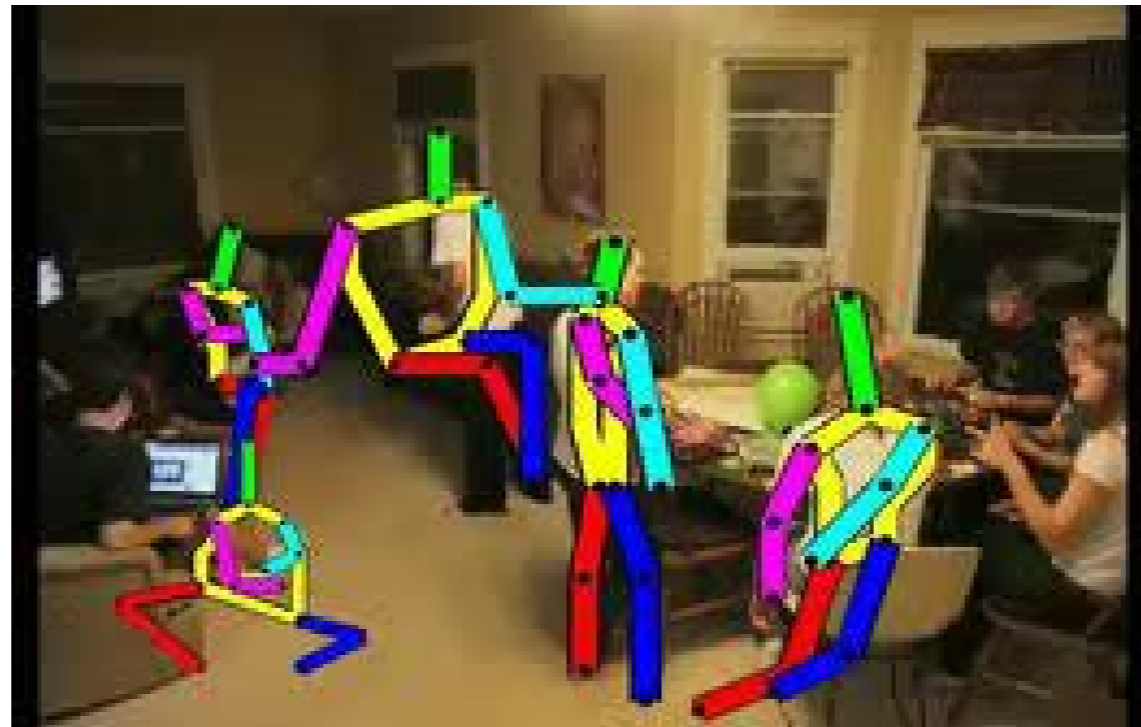


Output:  
3D Understanding

# Our Approach



Timelapse



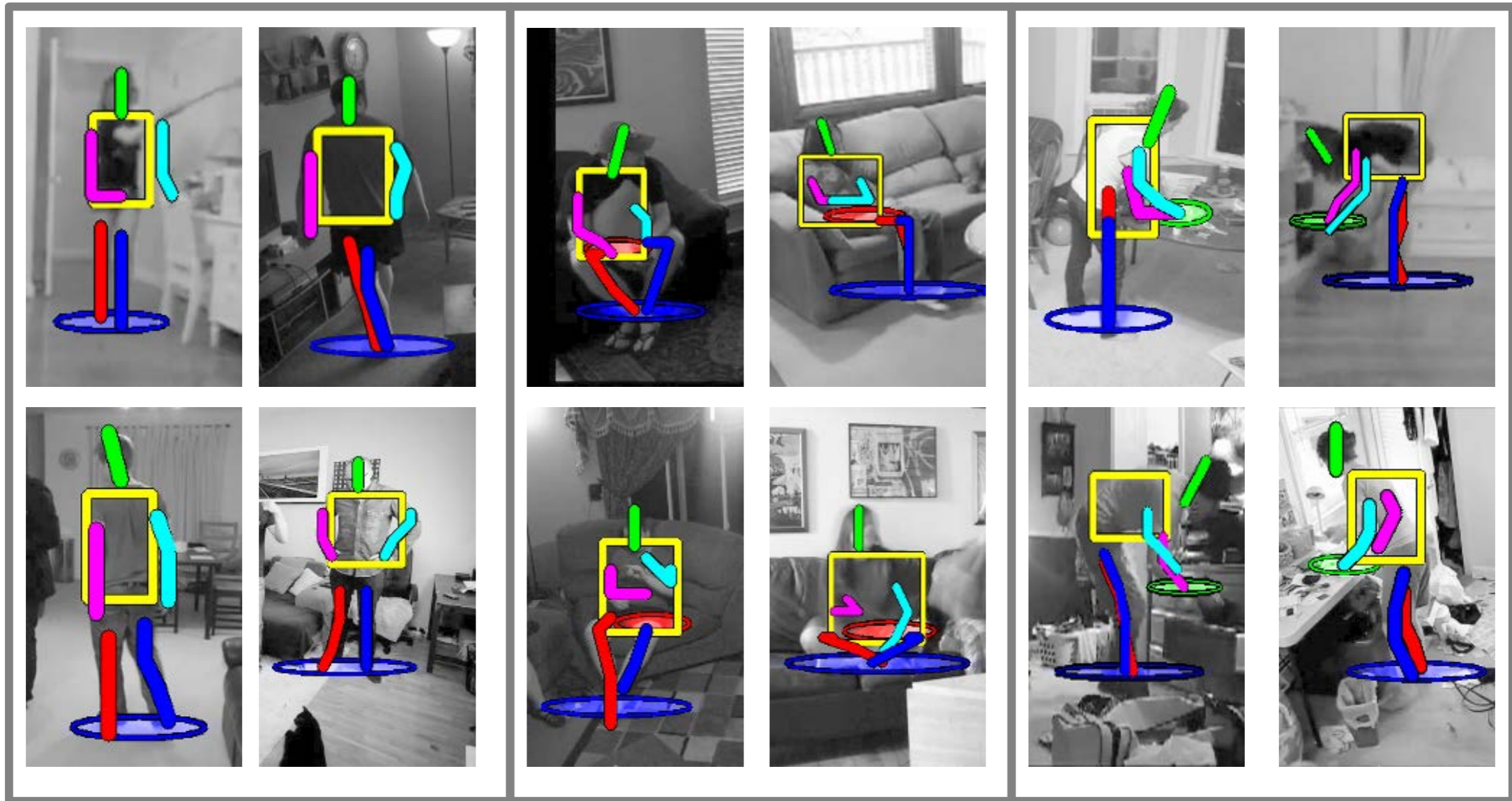
Pose Detections

# Detecting Human Actions

Standing

Sitting

Reaching



Yang and Ramanan '11

Train Separate Detectors for Each Pose

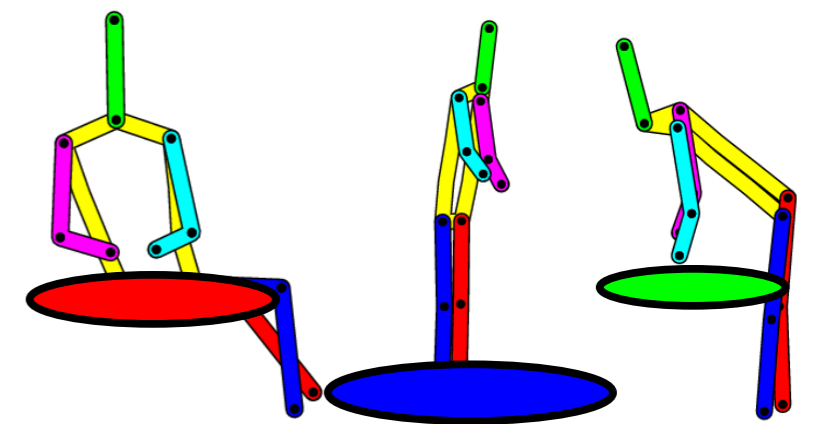
# Our Approach



Timelapse



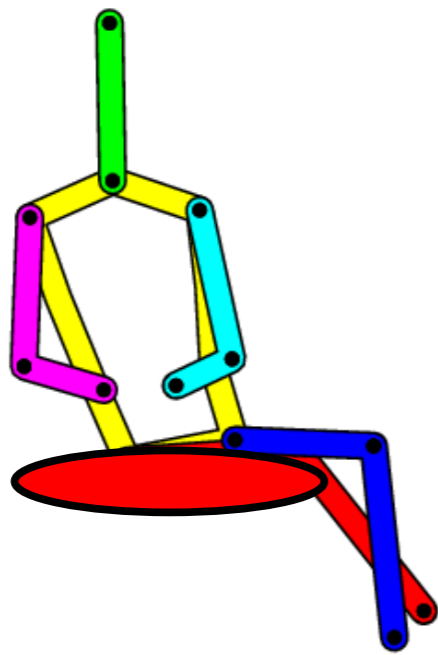
Pose Detections



Estimate Functional Regions from Poses

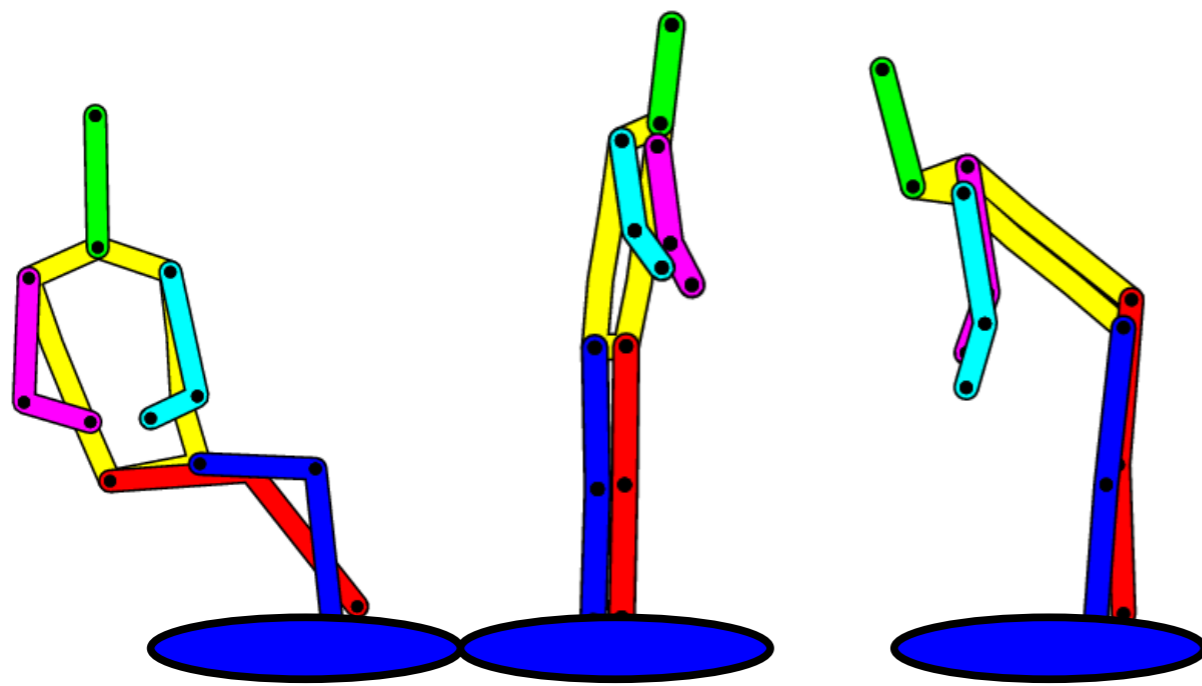


# From Poses to Functional Regions



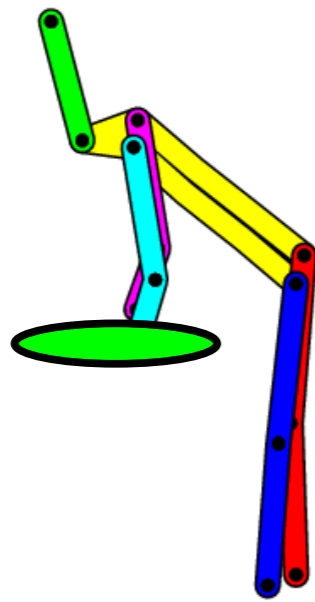
Sittable Regions at Pelvic Joint

# From Poses to Functional Regions



Walkable Regions at Feet

# Affordance Constraints



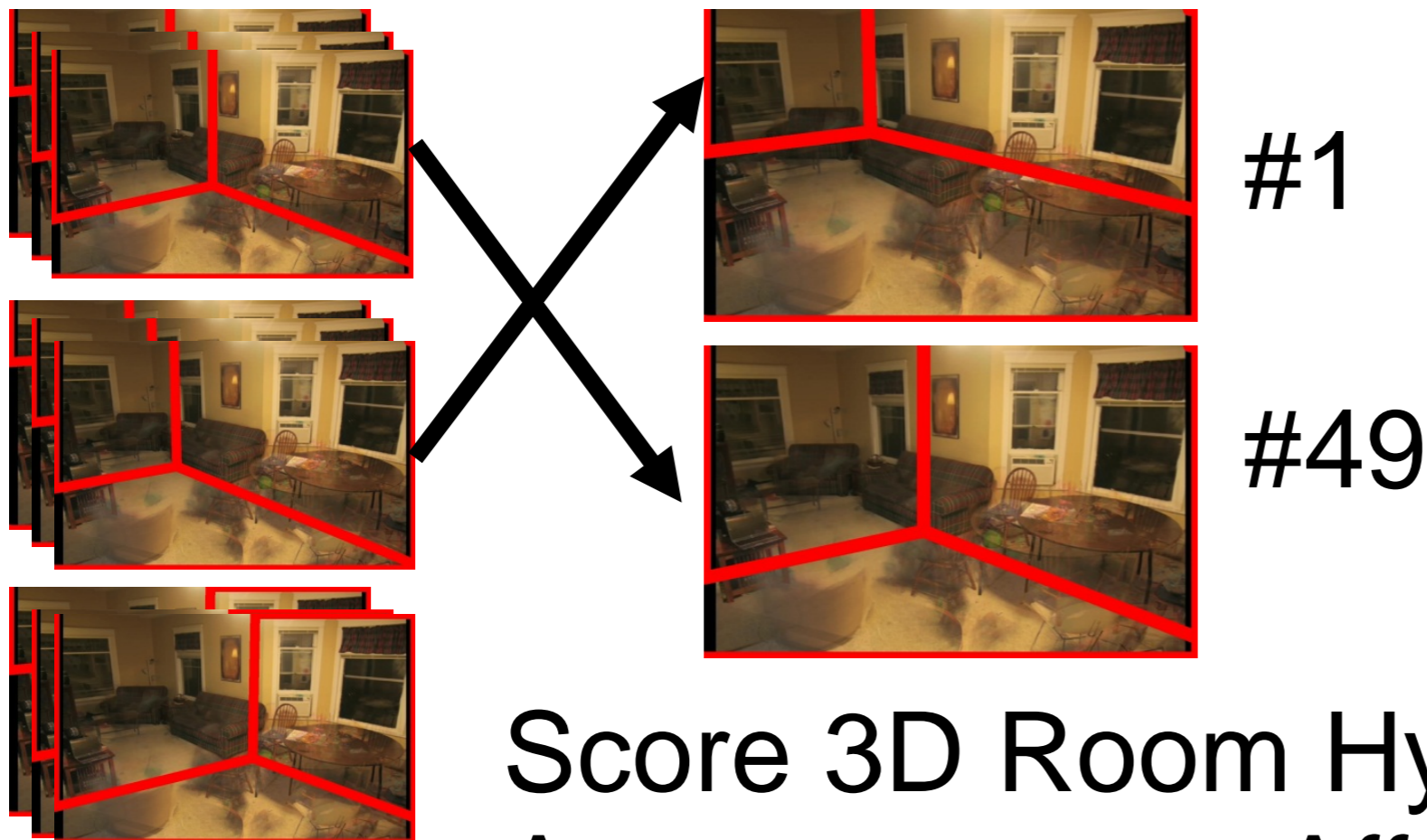
Reachable Regions at Hands

# Our Approach



3D Room Hypotheses From Appearance

# Our Approach



Score 3D Room Hypotheses With Appearances + Affordances

# Our Approach



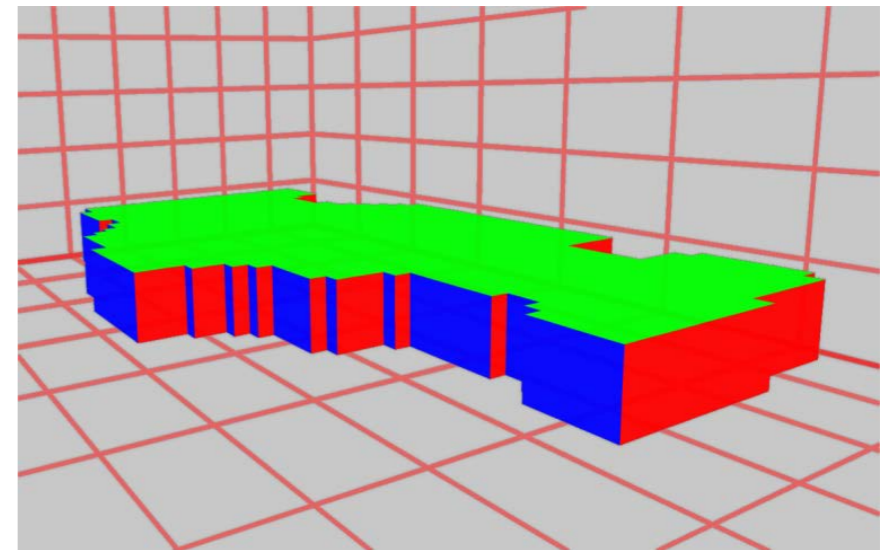
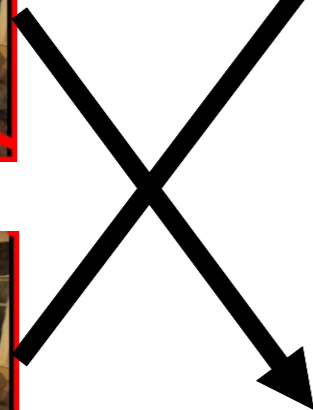
Timelapse



Pose Detections



Functional Regions

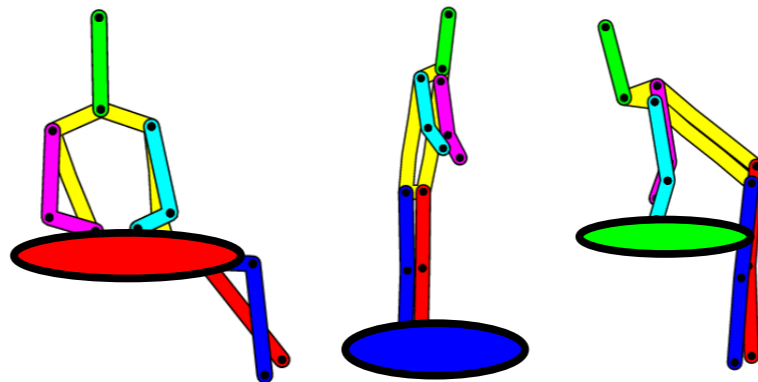


Estimate  
Free-Space

# Results

# Qualitative Example

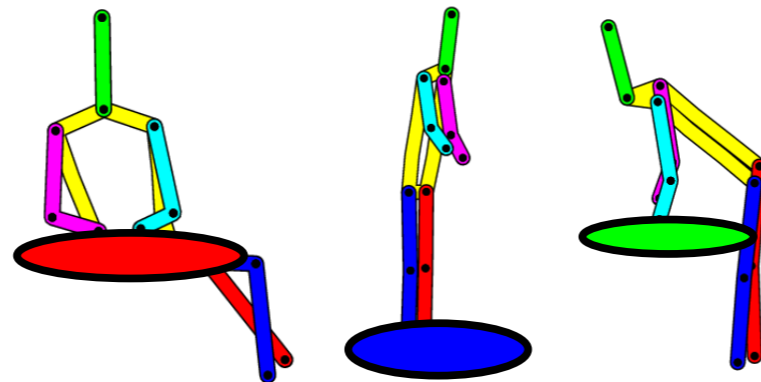
Original video





# Qualitative Example

Original video



# Quantitative Results

40 Timelapse videos from Youtube  
Evaluated on room layout estimation.

Location	Appearance Only		People Only	Appearance + People
	Lee et al. '09	Hedau et al. '09		
64.1%	70.4%	74.9%	70.8%	<b>82.5%</b>

Does equivalently or better 93% of the time



# Seeing 3D chairs:

## Exemplar part-based 2D-3D alignment using a large dataset of CAD models CVPR 2014

Mathieu Aubry (INRIA)

Daniel Maturana (CMU)

Alexei Efros (UC Berkeley)

Bryan Russell (Intel)

Josef Sivic (INRIA)

# Sit on the chair!

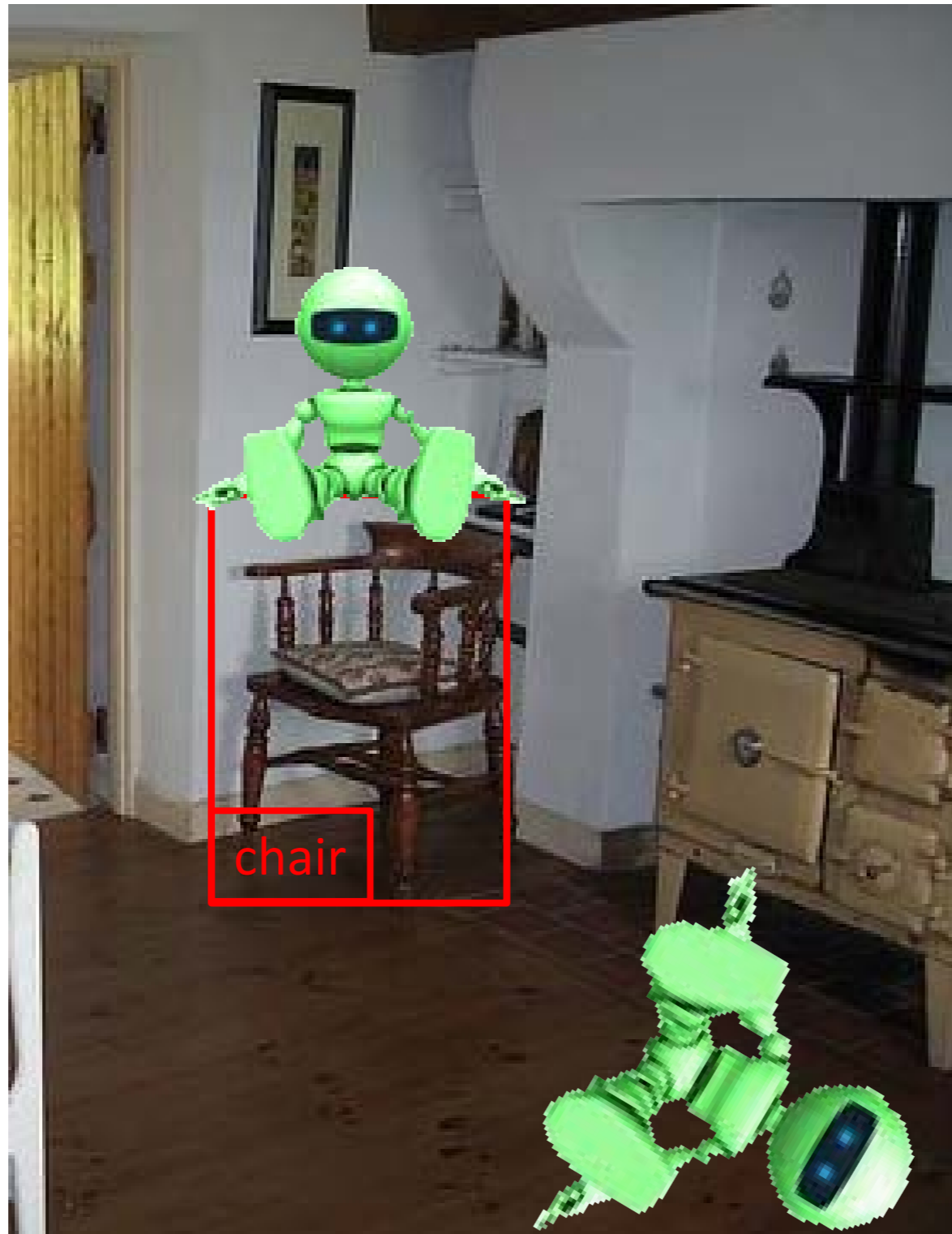


# Classification



Ex: ImageNet Challenge, Pascal VOC classification.

# Detection



Ex: Pascal VOC detection.

# Segmentation



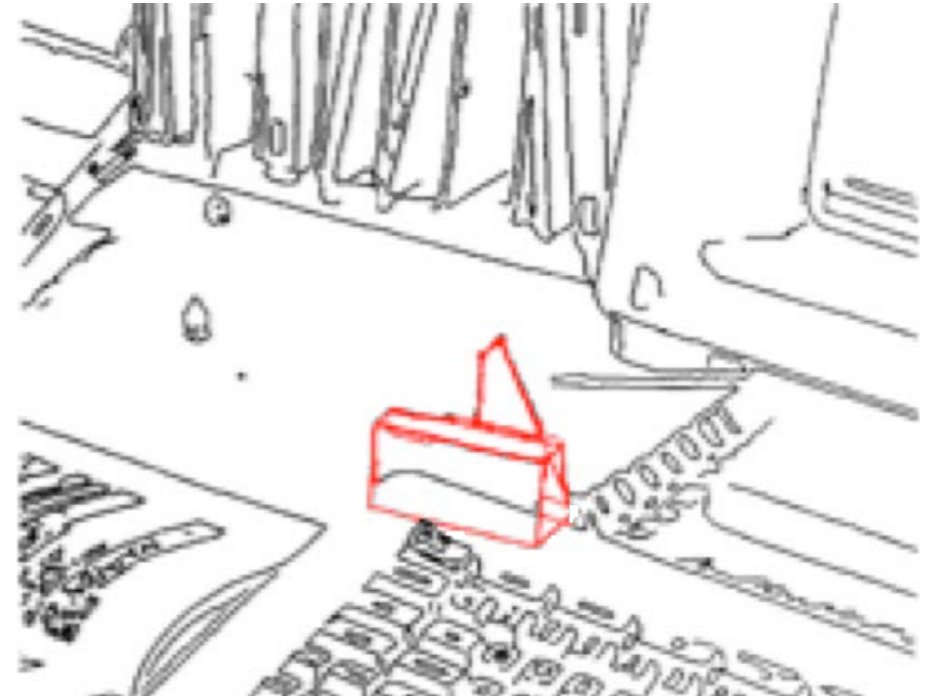
Ex: Pascal VOC segmentation.

# Our goal

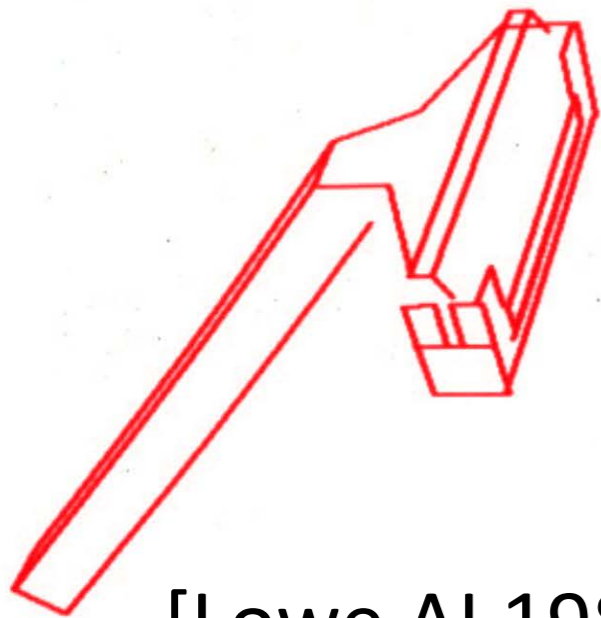




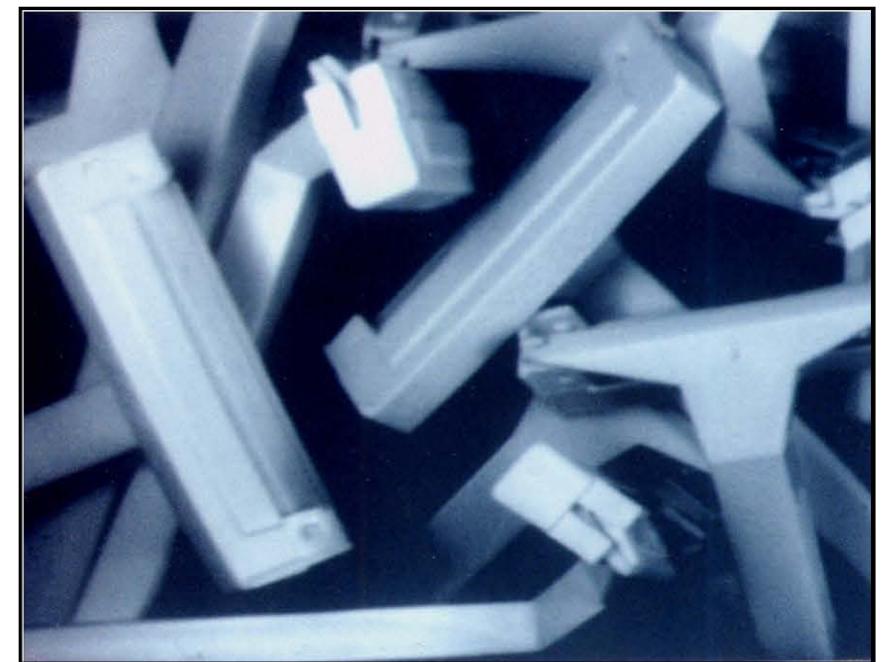
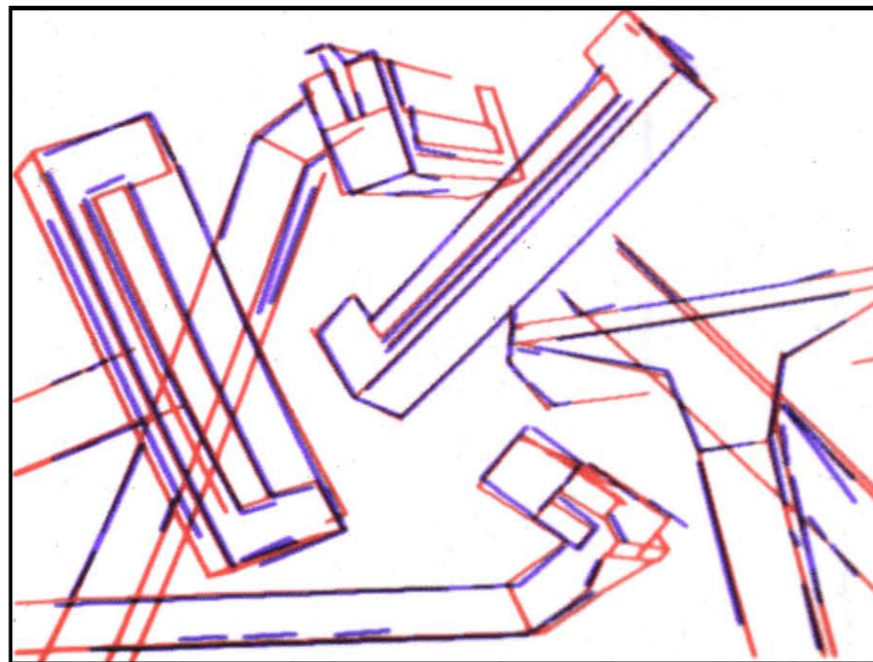
# 1980s: 2D-3D Instance Alignment



[Huttenlocher and Ullman IJCV 1990]

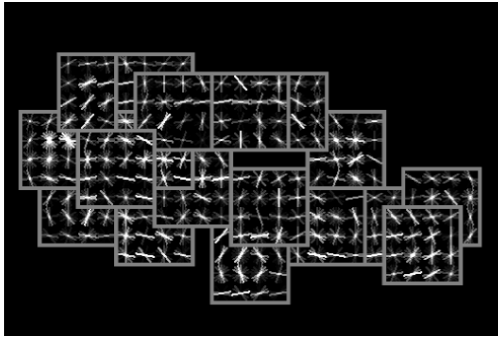


[Lowe AI 1987]



[Faugeras&Hebert'86], [Grimson&Lozano-Perez'86], ...

# Recent: 3D category recognition



3D DPMs: [Herjati&Ramanan'12], [Pepik et al.12], [Zia et al.'13], ...



Simplified part models: [Xiang&Savarese'12], [Del Pero et al.'13]



Cuboids: [Xiao et al.'12] [Fidler et al.'12]



Blocks world revisited: [Gupta et al.'12]

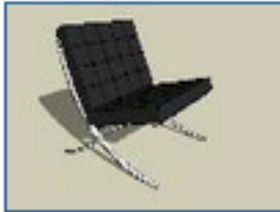








See also: [Glasner et al.'11], [Fouhey et al.'13], [Satkin&Hebert'13], [Choi et al. '13], [Hejrati and Ramanan '14], [Savarese and Fei-Fei '07]...

# Approach: data-driven

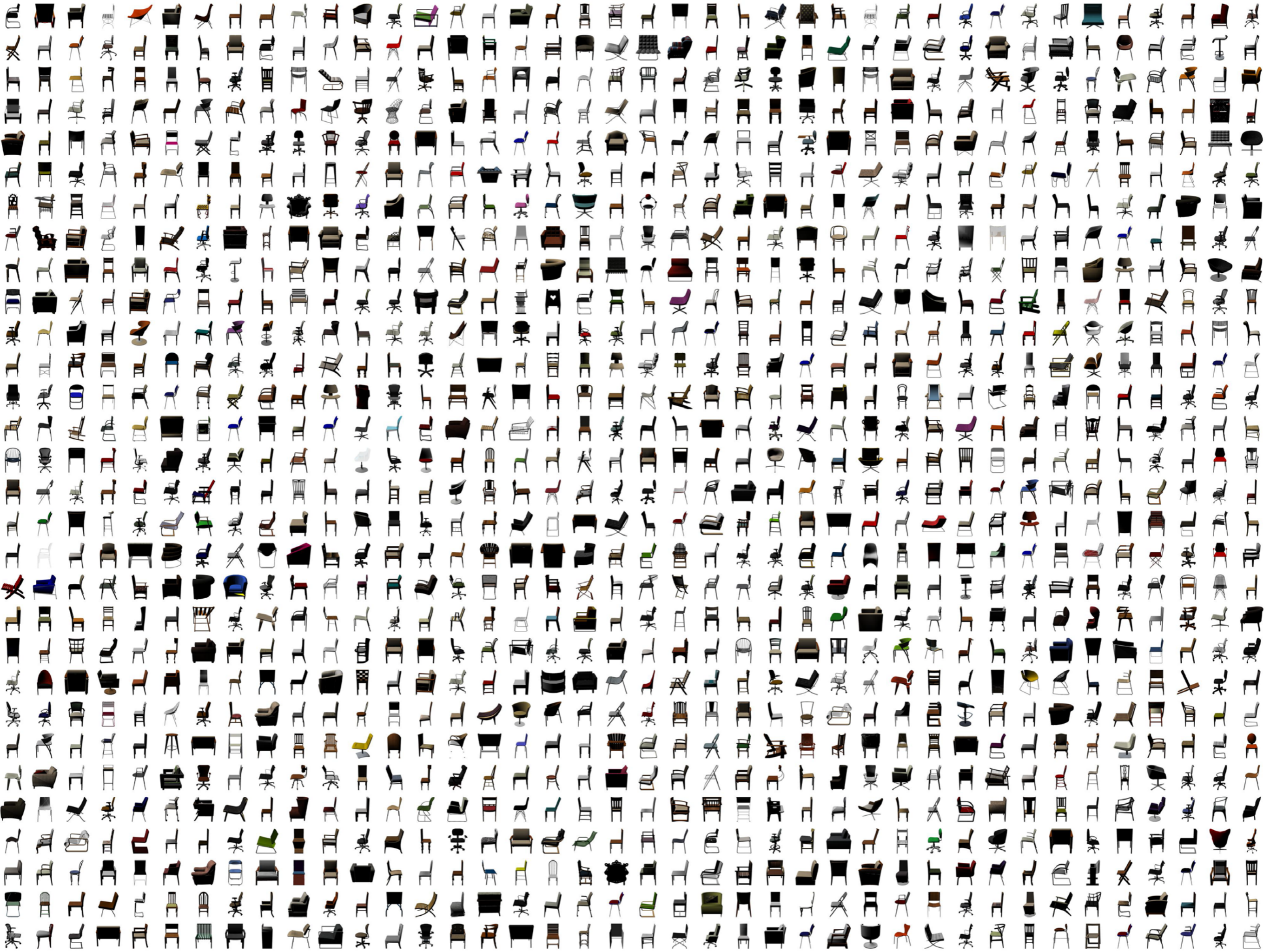
**Trimble 3D Warehouse** powered by Google

chair   [Advanced Search](#)

**3D Warehouse Results**

 ★★★★★	<b>Chair - Barcelona Chair</b> by <a href="#">Adam</a> Barcelona <b>Chair</b> 1929 Ludwig... <a href="#">Download to SketchUp 5</a>	 ★★★★★	<b>Eames Lounge Chair</b> by <a href="#">Mart</a> Lounge <b>Chair</b> and Ottoman by... <a href="#">Download to SketchUp 6</a>	 ★★★★★	<b>LC-2 (Chair)</b> by <a href="#">Archi Maniac</a> One of the most famous modern... <a href="#">Download to SketchUp 5</a>
 ★★★★★	<b>Barcelona Chair</b> by <a href="#">acad whiz</a> EDITED: Added the support... <a href="#">History</a> <a href="#">Download to SketchUp 6</a>	 ★★★★★	<b>Egg Chair</b> by <a href="#">Mart</a> Egg <b>Chair</b> by Arne Jacobsen... <a href="#">Download to SketchUp 6</a>	 ★★★★★	<b>Eames Lounge Chair &amp; Ottoman...</b> by <a href="#">SmartFurniture.com</a> The Eames Lounge <b>Chair</b> and... <a href="#">History</a> <a href="#">Download to SketchUp 8</a>
 ★★★★★	<b>chair silla</b> by <a href="#">ketchup</a> silla clasica-----classic... <a href="#">Download to SketchUp 6</a>	 ★★★★★	<b>BKF Chair and the BKF 2000</b> by <a href="#">ArgDirk</a> The BKF 2000 <b>chair</b> or bench... <a href="#">Download to SketchUp 6</a>	 ★★★★★	<b>Wassily Chair, Model B3 chair</b> by <a href="#">Darrell Smith</a> Deducing Moon's import Stuhl... <a href="#">Download to SketchUp 6</a>

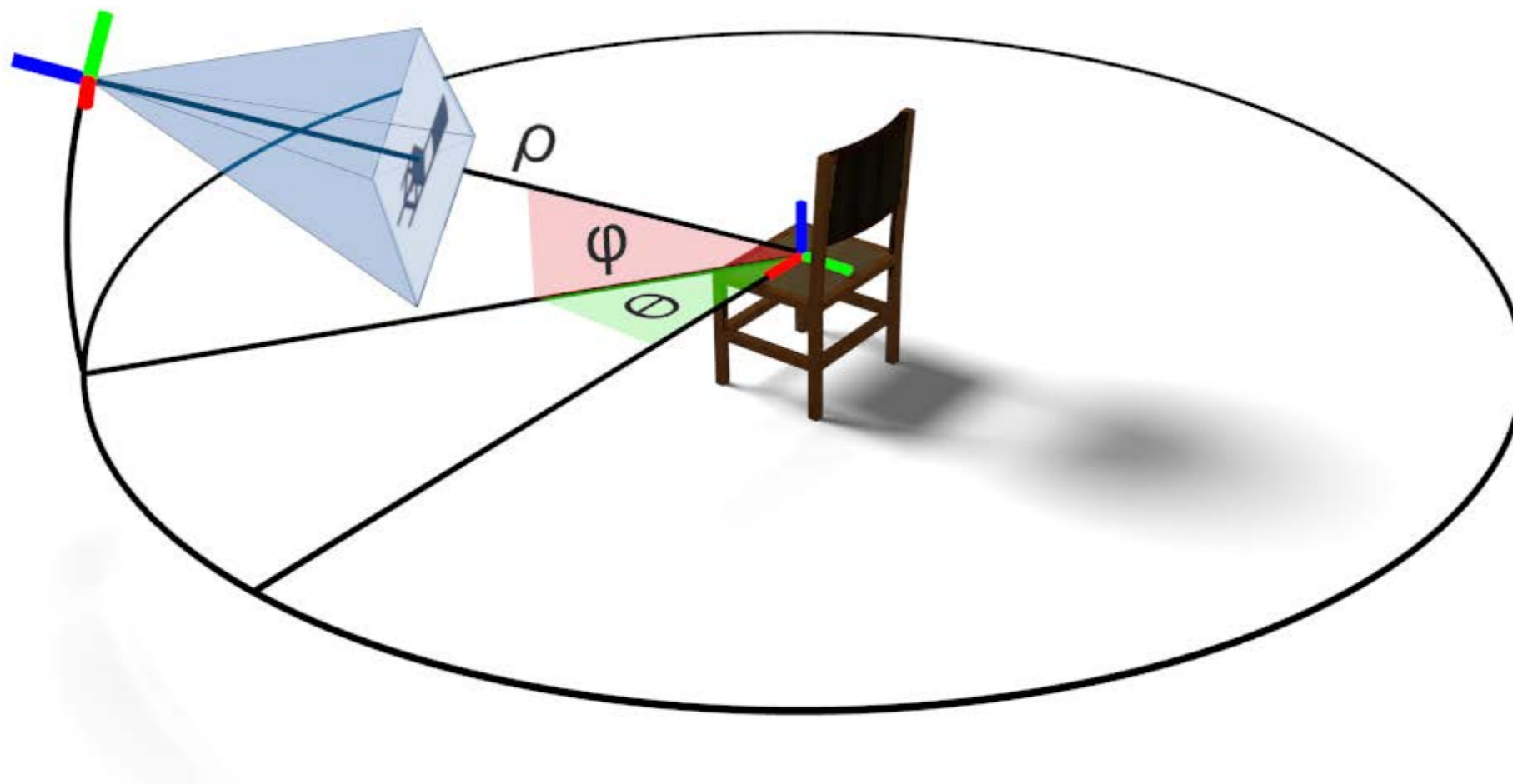
1394 3D models from internet



# Difficulty: viewpoint



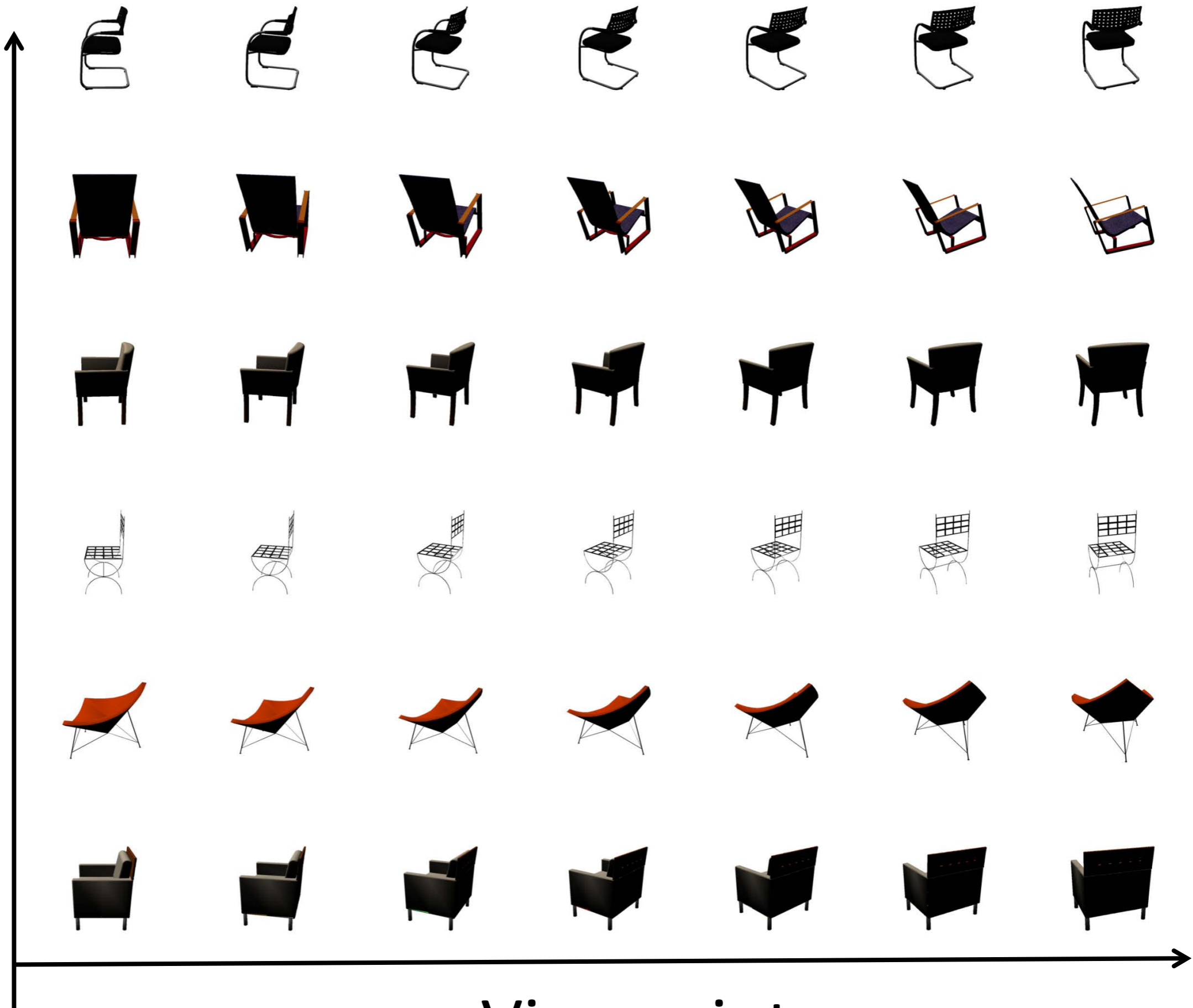
# Approach: use 3D models



62 views



Style



Viewpoint

# Difficulty: approximate style





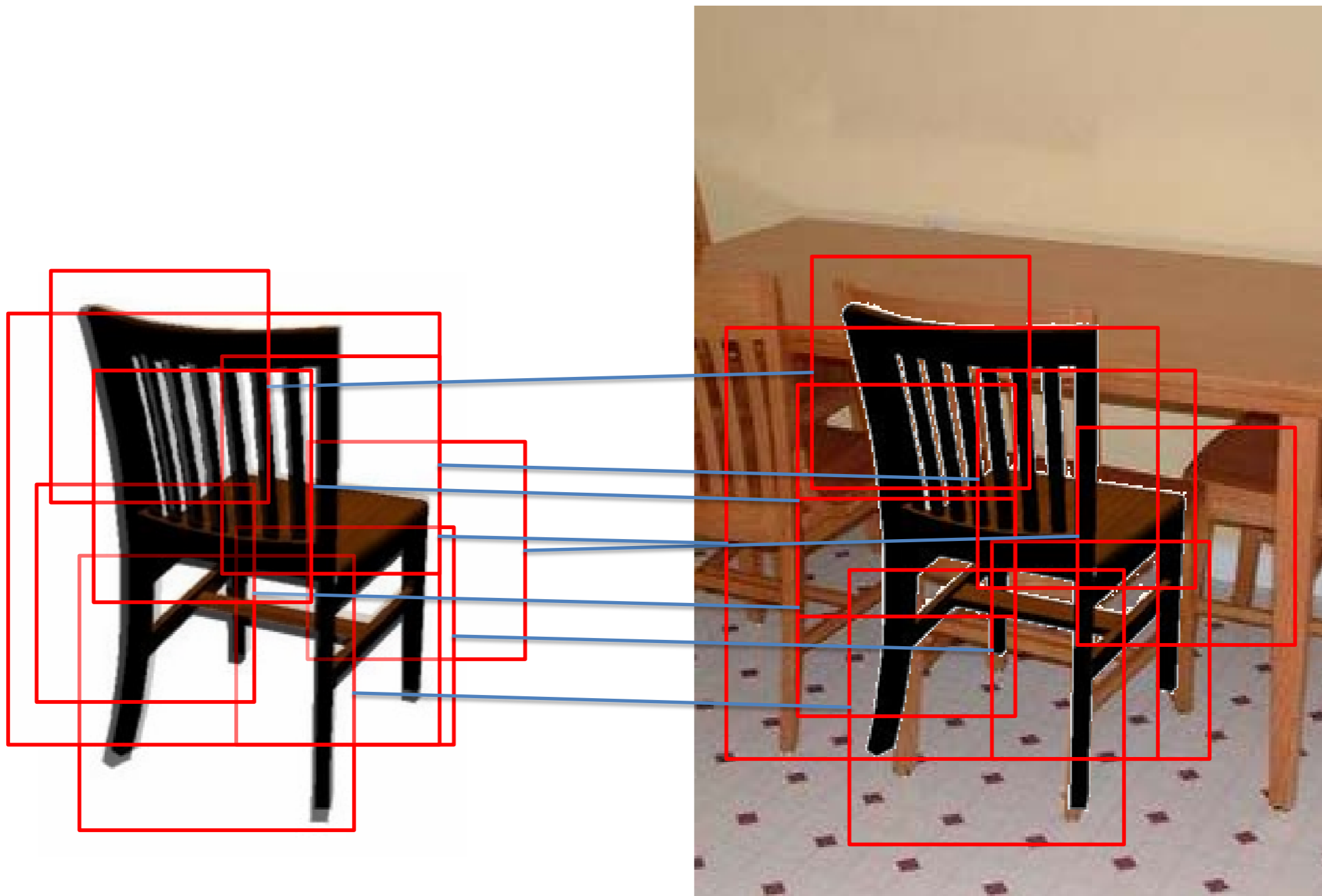
# Difficulty: approximate style



# Difficulty: approximate style



# Approach: part-based model

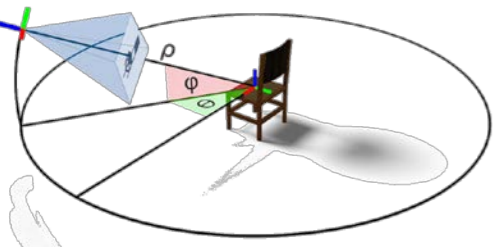


# Approach overview

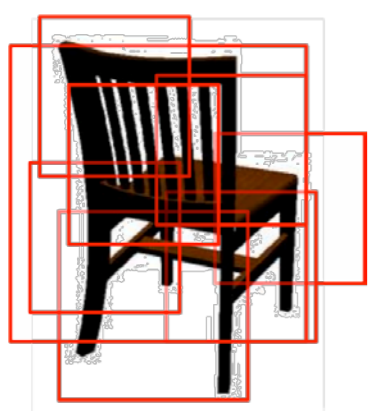
3D collection



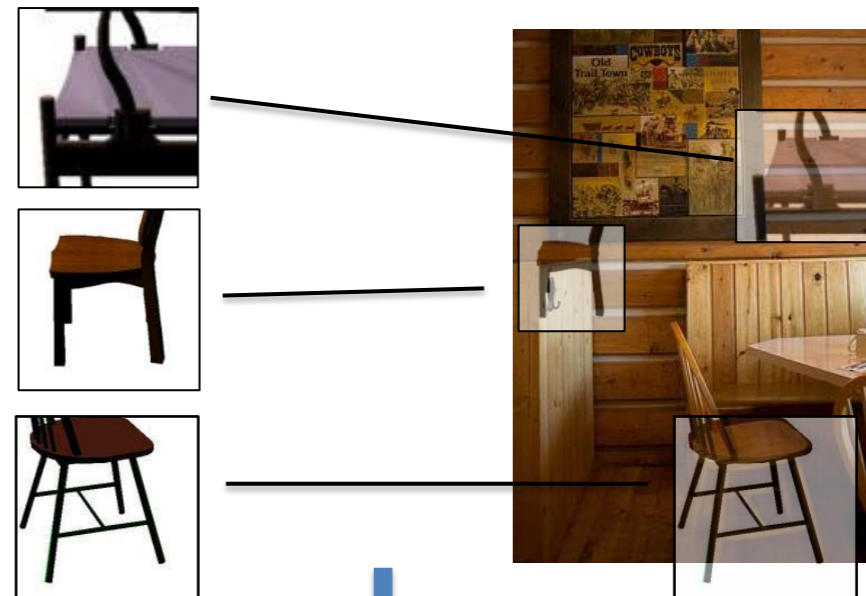
Render views



Select parts

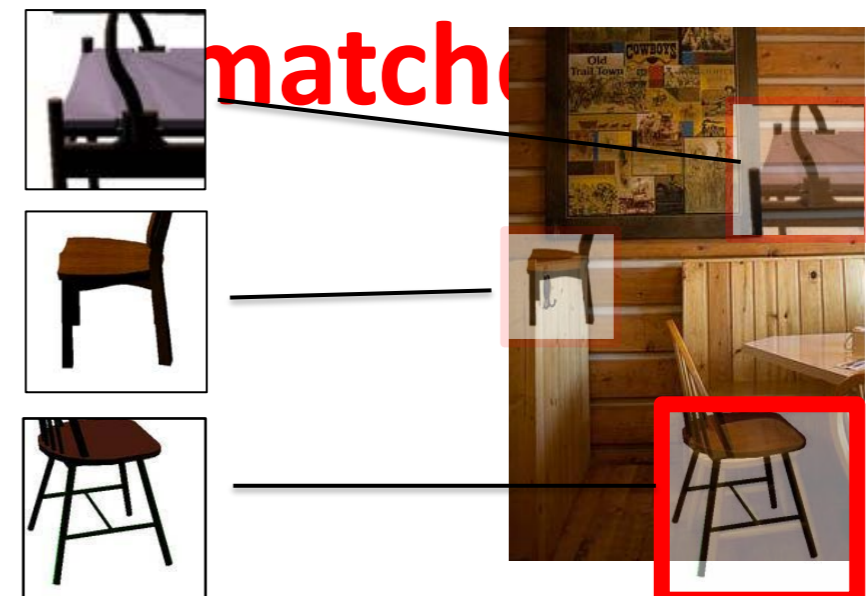


Match CG->real image

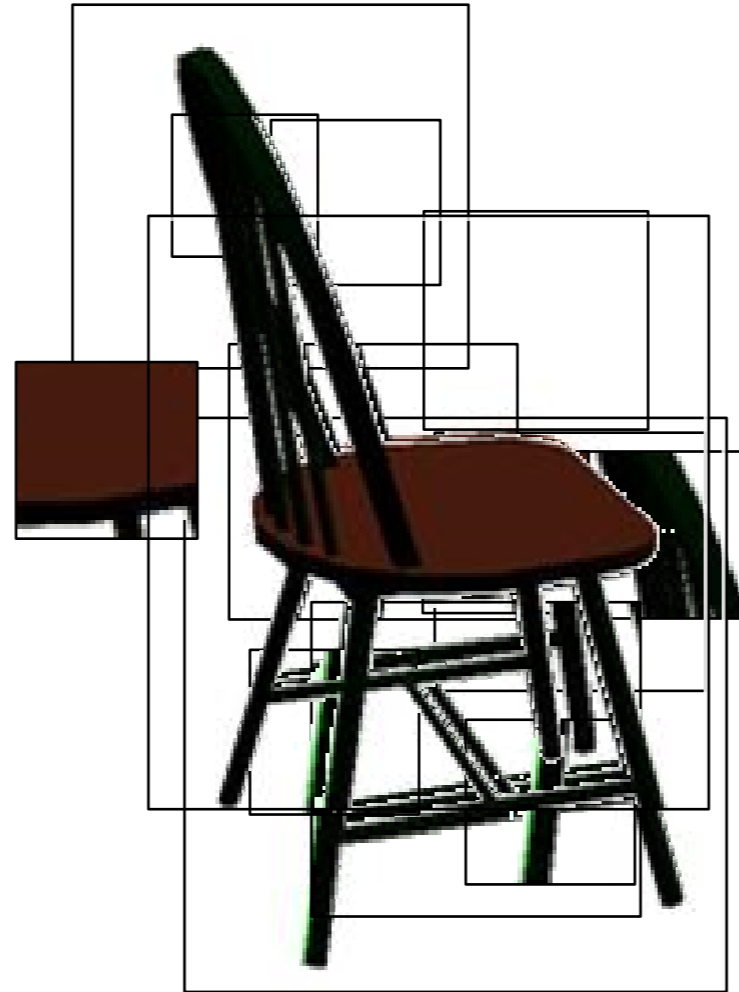
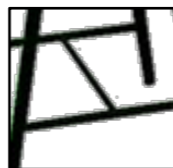


Select the best

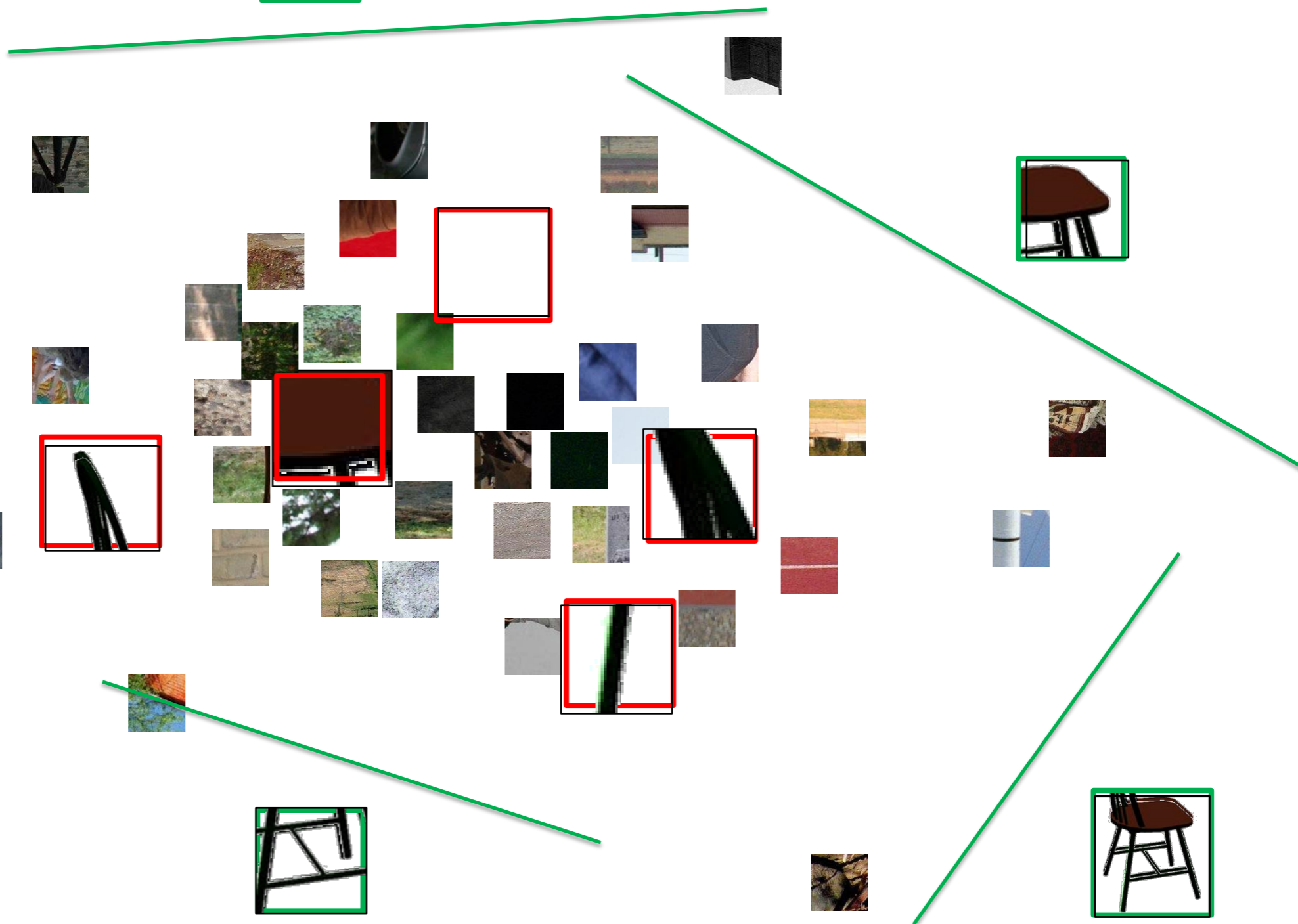
match



# Select discriminative parts



# How to select discriminative parts?



Best exemplar-LDA classifiers

[Hariharan et al. 2012] [Gharbi et al 2012]  
[Malisiewicz et al 2011]

# Approach: CG-to-photograph



Implementation: exemplar-LDA

# How to compare matches?

Patches



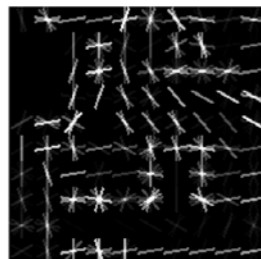
Detectors



$$(w_1, b_1)$$



$$(w_2, b_2)$$



$$(w_3, b_3)$$

Matches



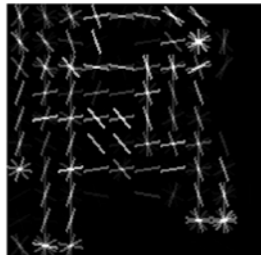


# How to compare matches?

Patches



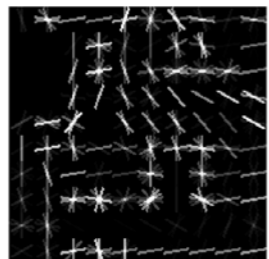
Detectors



$$(w_1, b_1)$$

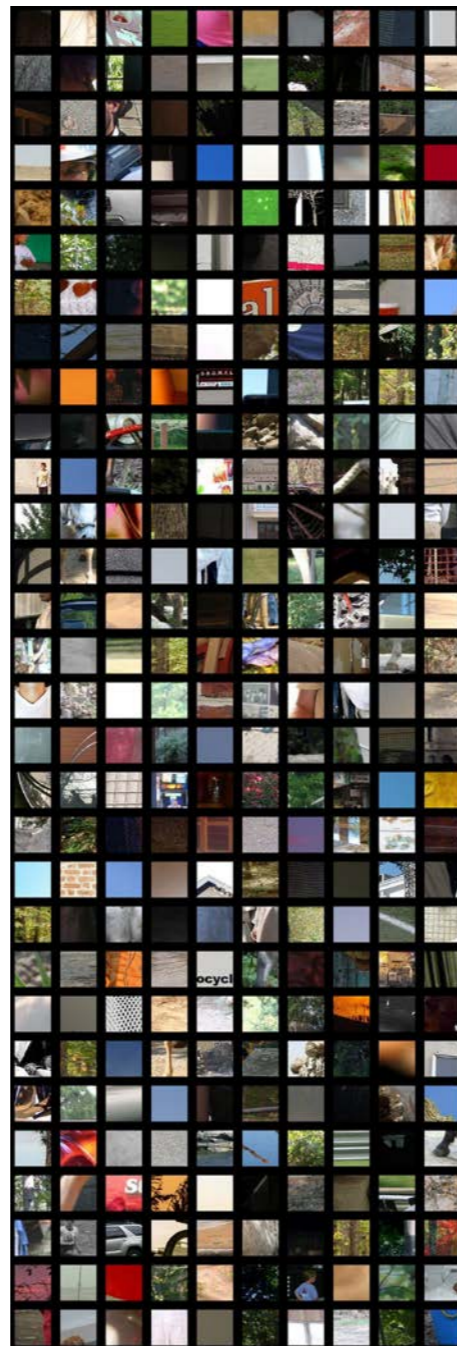


$$(w_2, b_2)$$



$$(w_3, b_3)$$

Affine Calibration  
with negative data



$$(\underline{a_1 w_1}, \underline{b'_1})$$

$$(\underline{a_2 w_2}, \underline{b'_2})$$

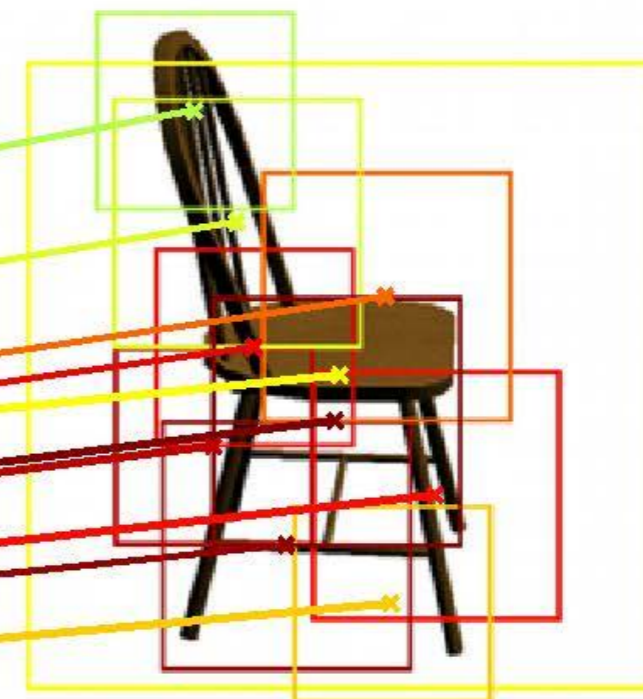
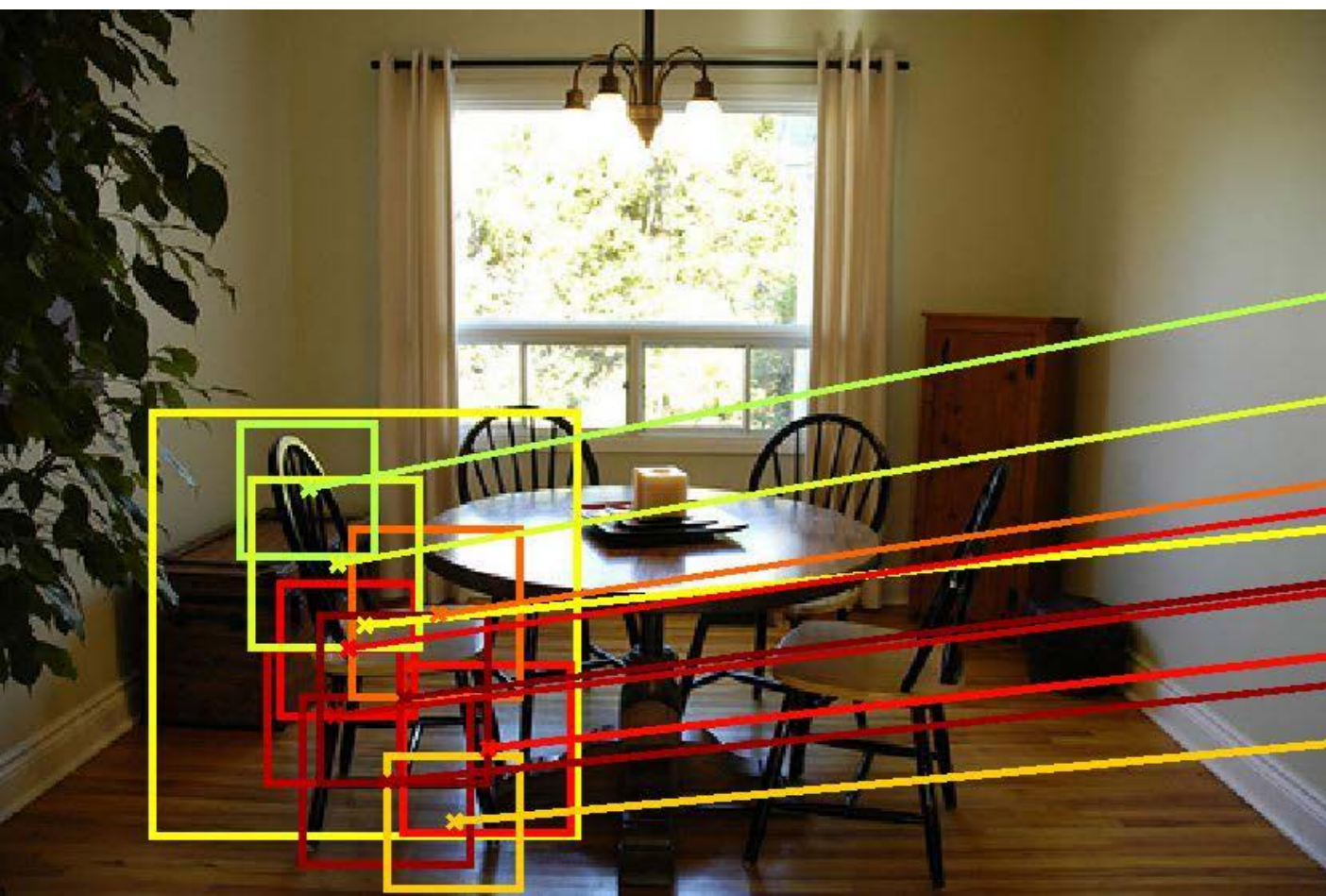
$$(\underline{a_3 w_3}, \underline{b'_3})$$

See paper for details

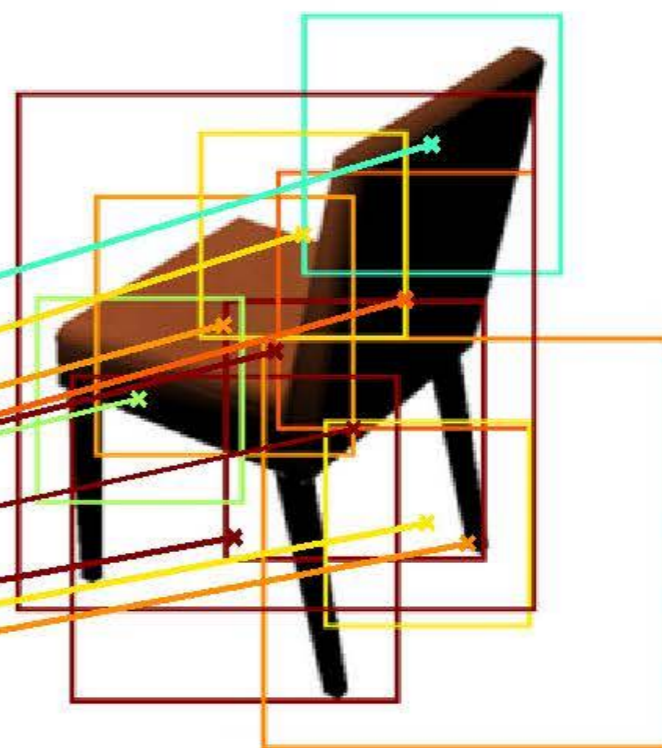
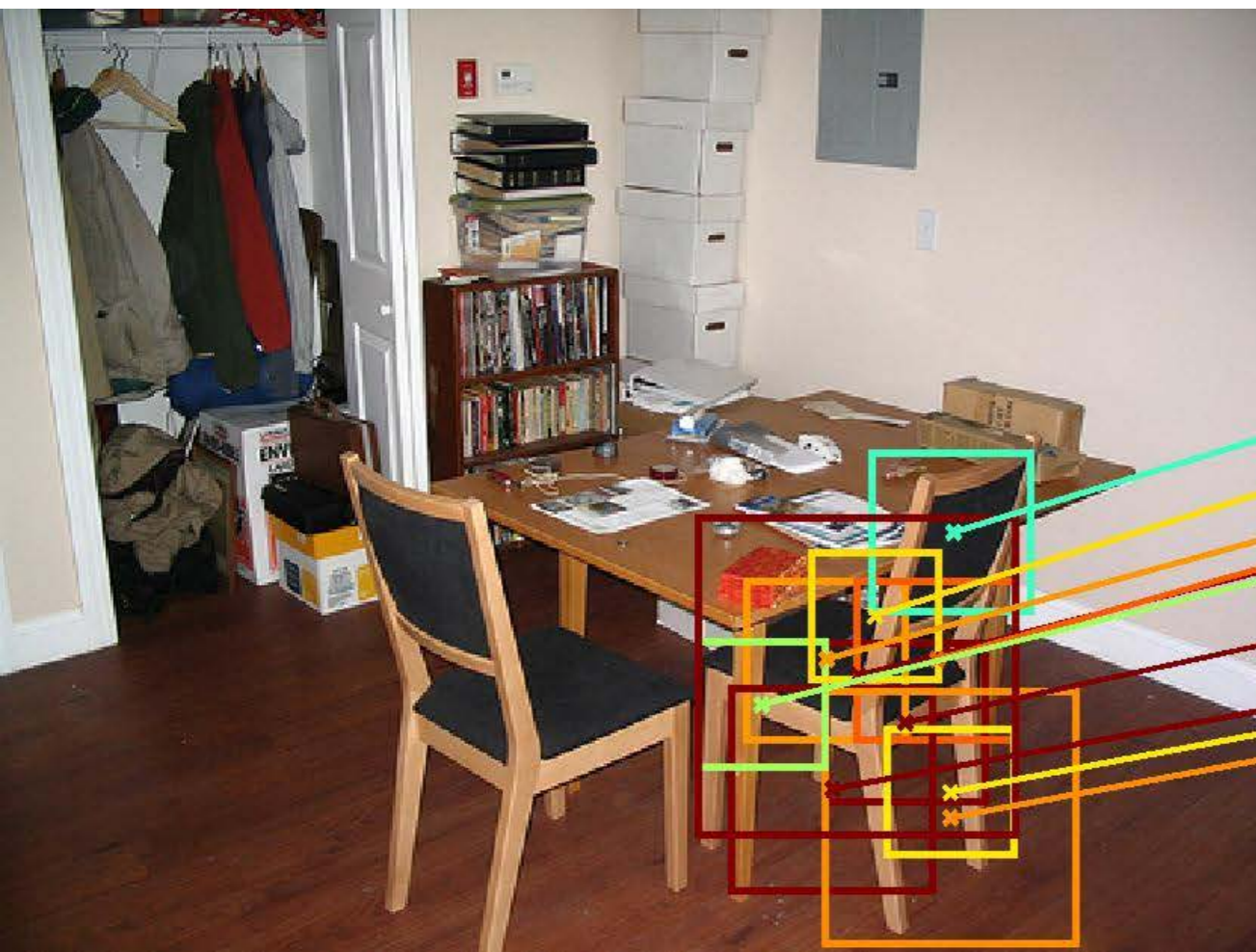
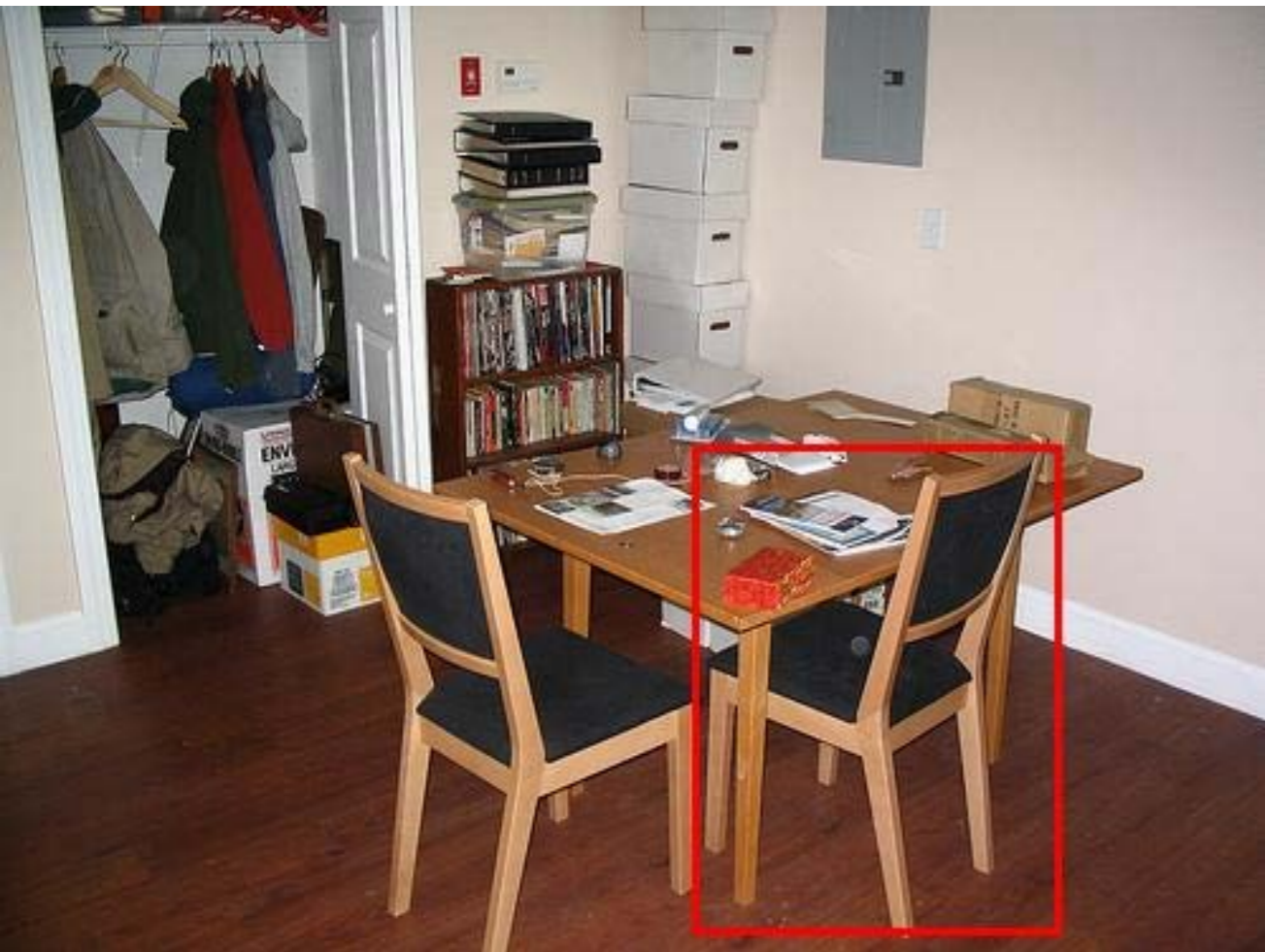
Matches



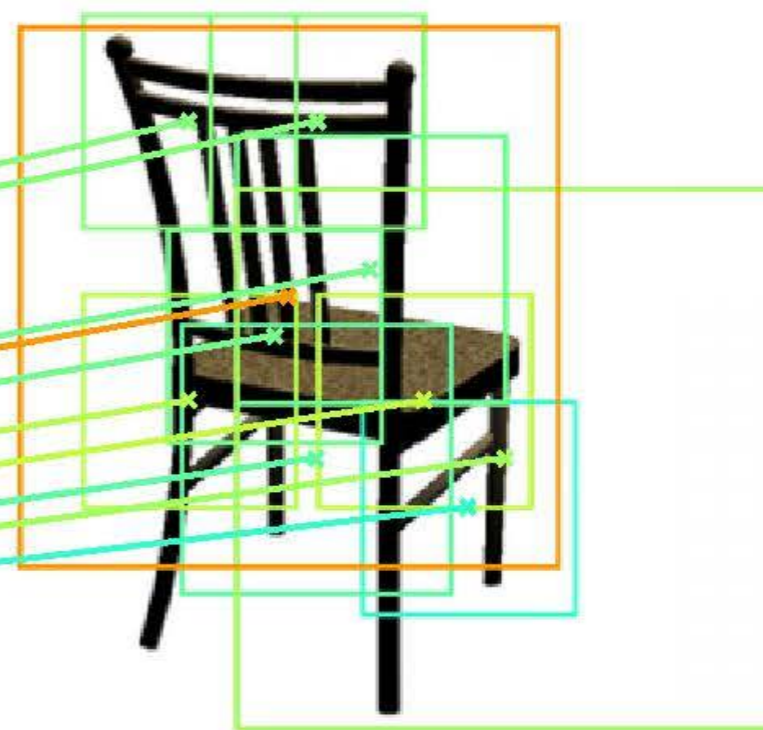
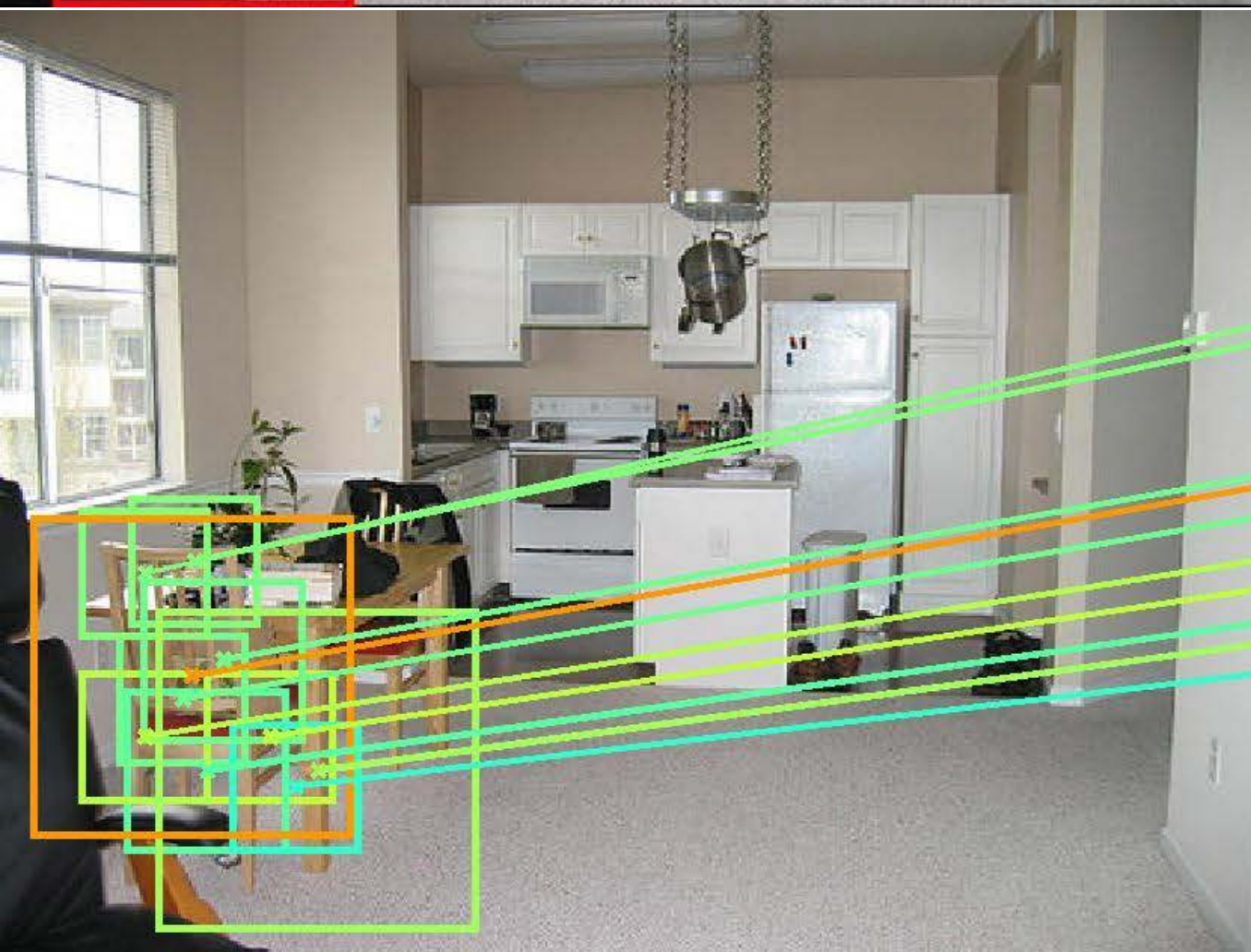
Example 1.



Example II.

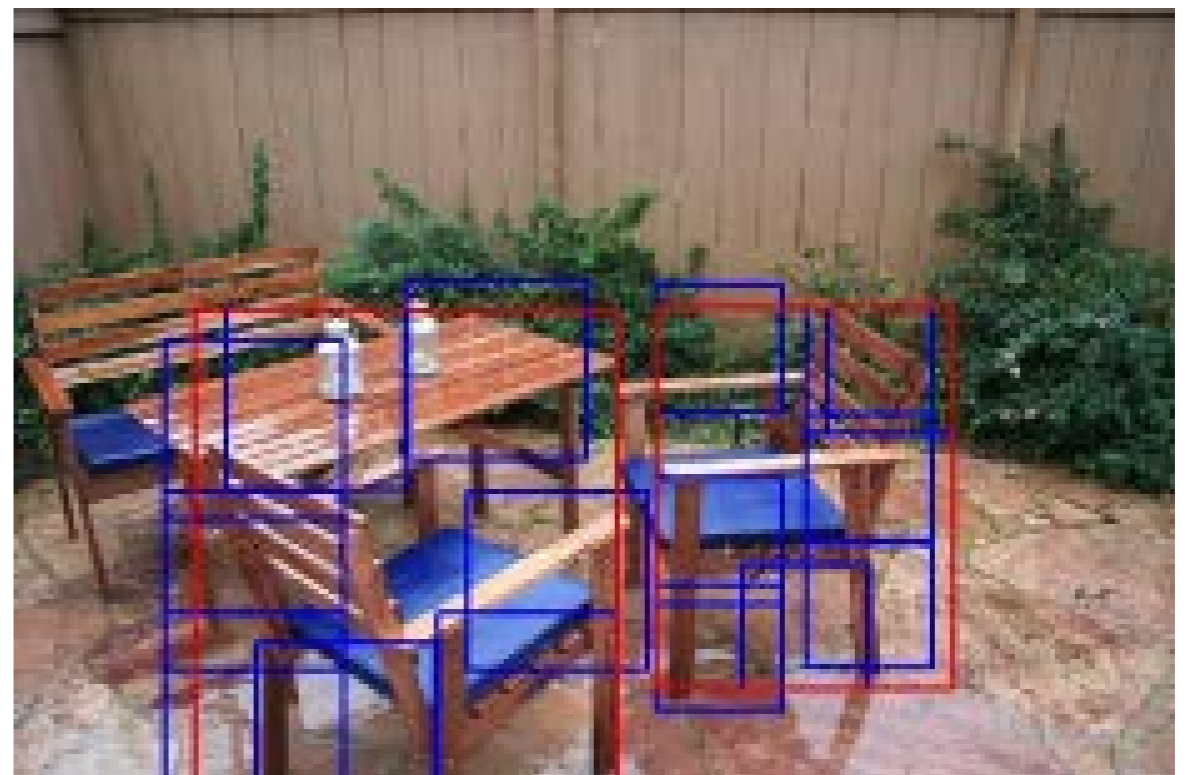


Example III.





Input image



DPM output



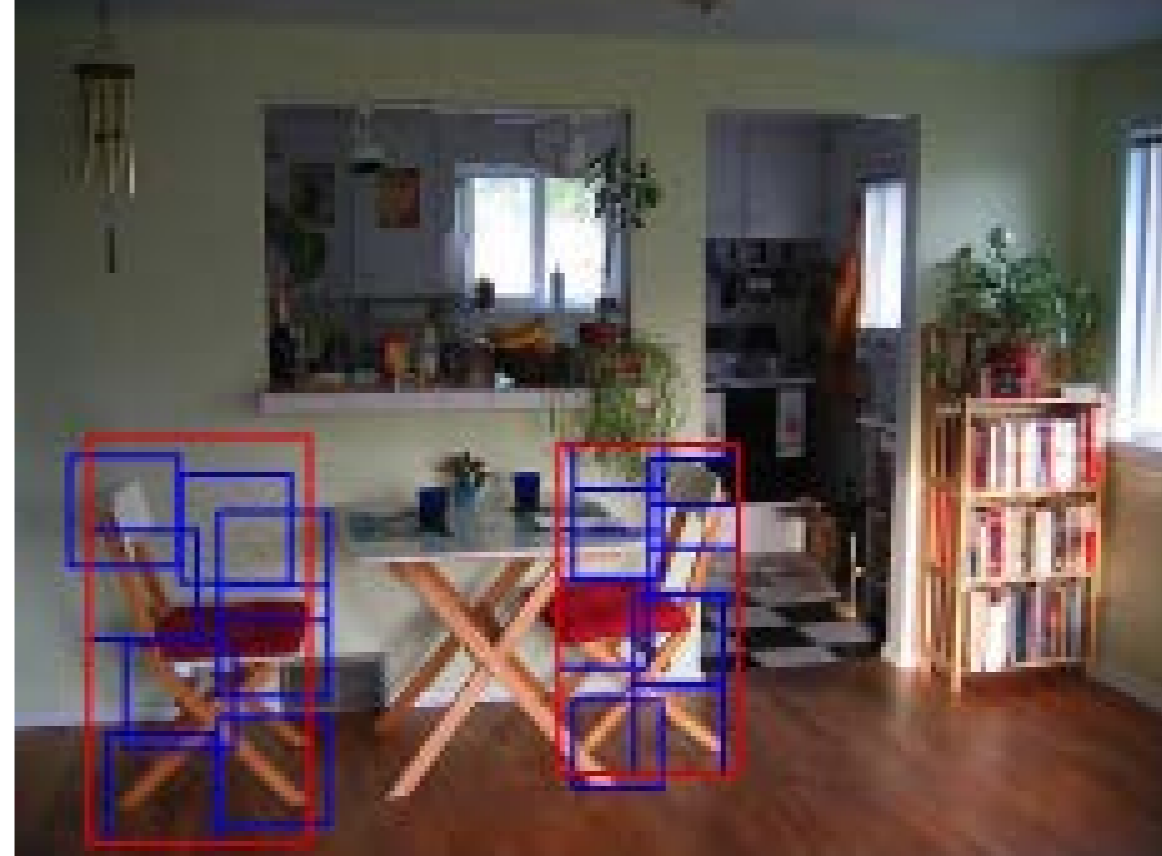
Our output



3D models



Input image



DPM output



Our output



3D models

# human evaluation

## Orientation quality at 25% recall



	Good	Bad
Exemplar-LDA	52%	48%
Ours	<b>90%</b>	<b>10%</b>

# human evaluation

## Style consistency at 25% recall



	Exact	Ok	Bad
Exemplar-LDA	3%	31%	66%
<b>Ours</b>	<b>21%</b>	<b>64%</b>	<b>15%</b>





# 3D Object Manipulation in a Single Photograph using Stock 3D Models

Natasha Kholgade<sup>1</sup>

Tomas Simon<sup>1</sup>

Alexei Efros<sup>2</sup>

Yaser Sheikh<sup>1</sup>

<sup>1</sup>Carnegie Mellon University

<sup>2</sup>University of California, Berkeley





# 3D Object Manipulation in a Single Photograph using Stock 3D Models

Natasha Kholgade<sup>1</sup>

Tomas Simon<sup>1</sup>

Alexei Efros<sup>2</sup>

Yaser Sheikh<sup>1</sup>

<sup>1</sup>Carnegie Mellon University

<sup>2</sup>University of California, Berkeley



Original Photograph



Object Manipulated in 3D



3D Copy-Paste

# Mental Picture



## The Language Bottleneck

**words**



# Image