Lecture 1

Introduction to computer vision

6.869/6.819 Advances in Computer Vision

Spring 2021

Bill Freeman, Phillip Isola

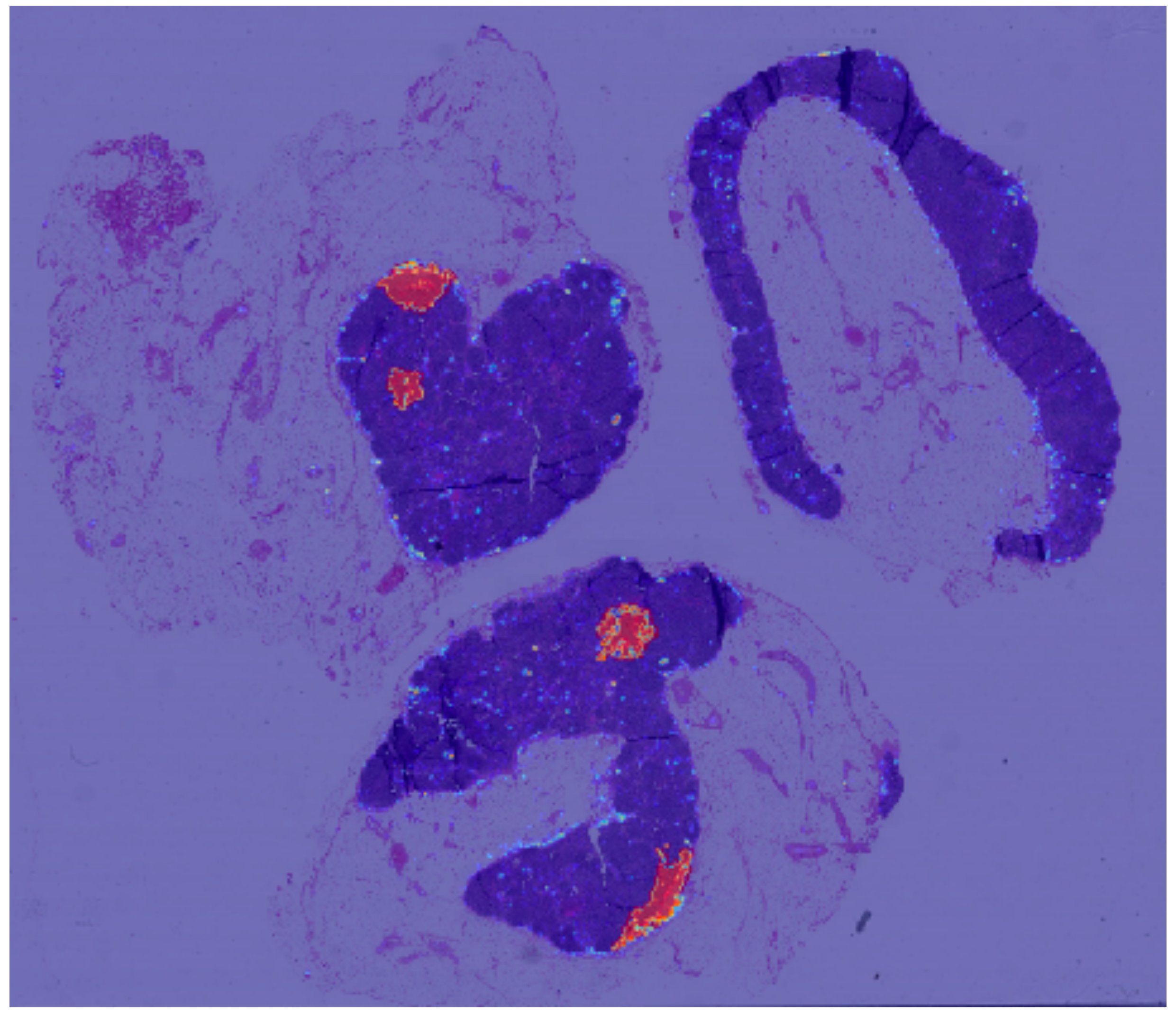
1. Introduction to computer vision

- History
- Perception versus measurement
- Simple vision system
- Taxonomy of computer vision tasks

Exciting times for computer vision

Robotics

Medical applications



Gaming

Driving

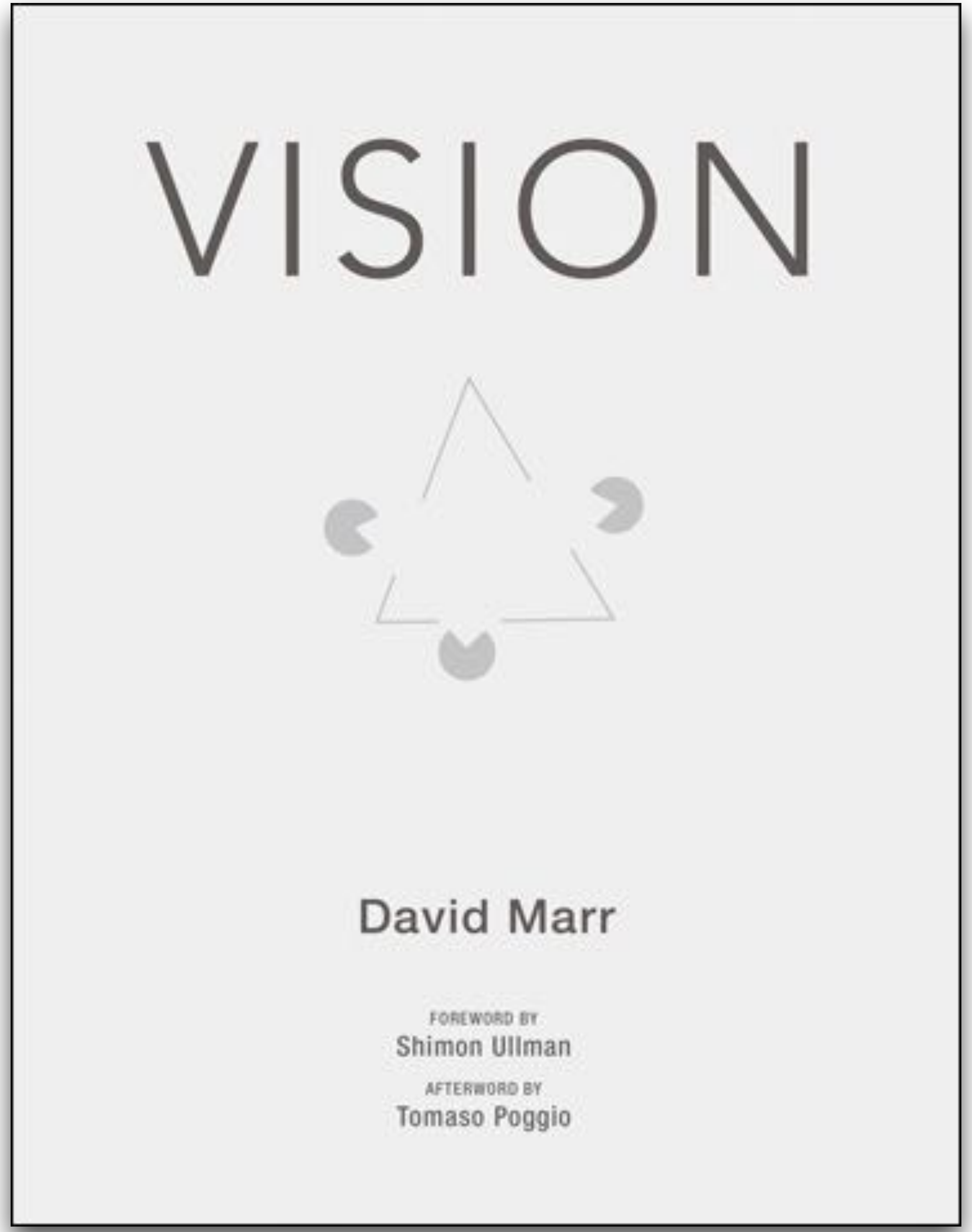
Mobile devices

Accessibility

To see

"What does it mean, to see? The plain man's answer (and Aristotle's, too). would be, to know what is where by looking."

To discover from images what is present in the world, where things are, what actions are taking place, to predict and anticipate events in the world.



DSpace@MIT

Search (Ex: crystalline silicon solar)

Search Within This Collection   Advanced Search

Home → Computer Science and Artificial Intelligence Lab (CSAIL) → Artificial Intelligence Lab Publications → AI Memos (1959 - 2004) → View Item

▾ Browse

All of DSpace@MIT
Communities & Collections
By Issue Date
Authors
Titles
Subjects

This Collection
By Issue Date
Authors
Titles
Subjects

▸ My Account

The Summer Vision Project

Download

Author: Papert, Seymour A.

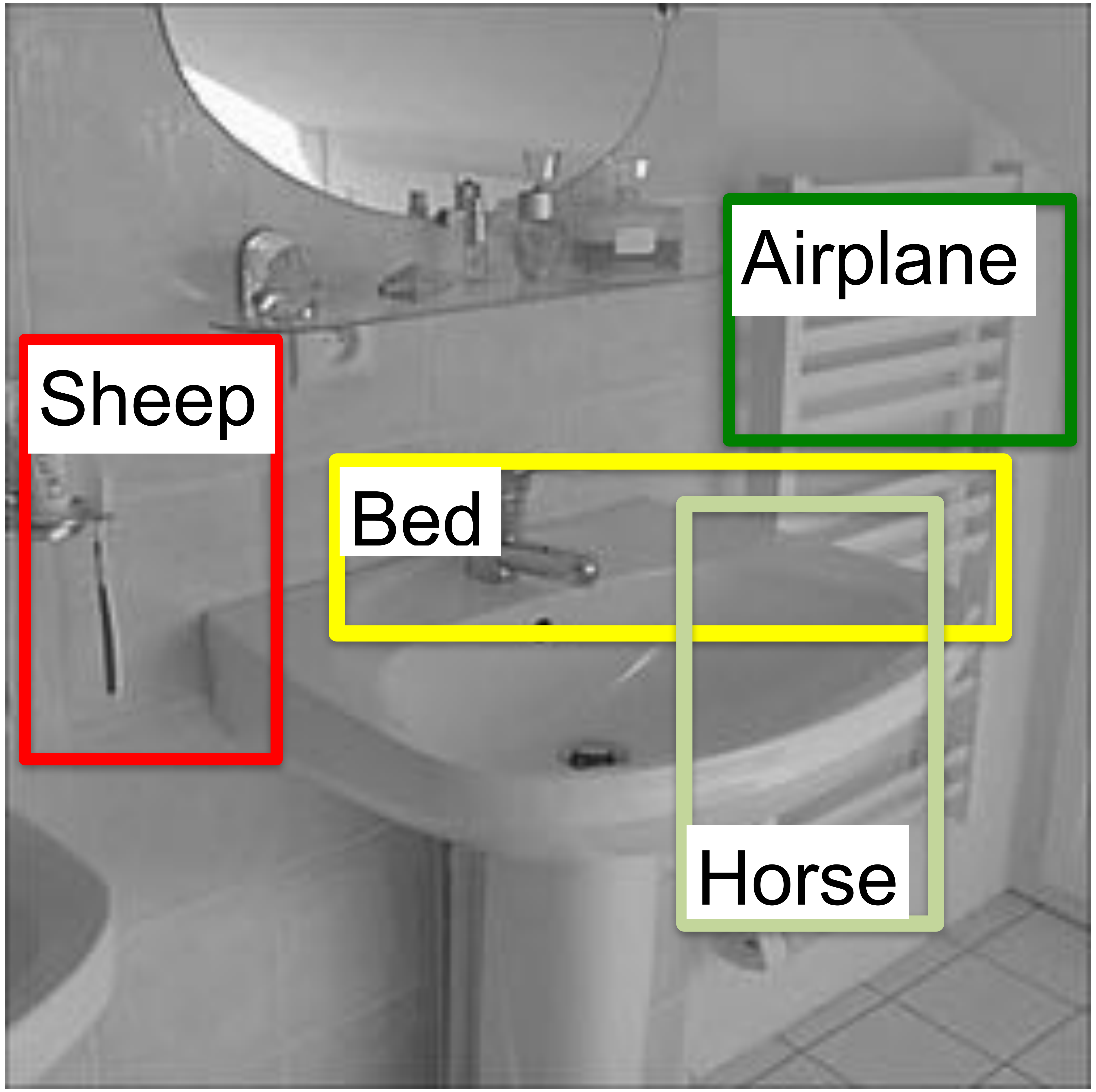
Citable URI: http://hdl.handle.net/1721.1/6125

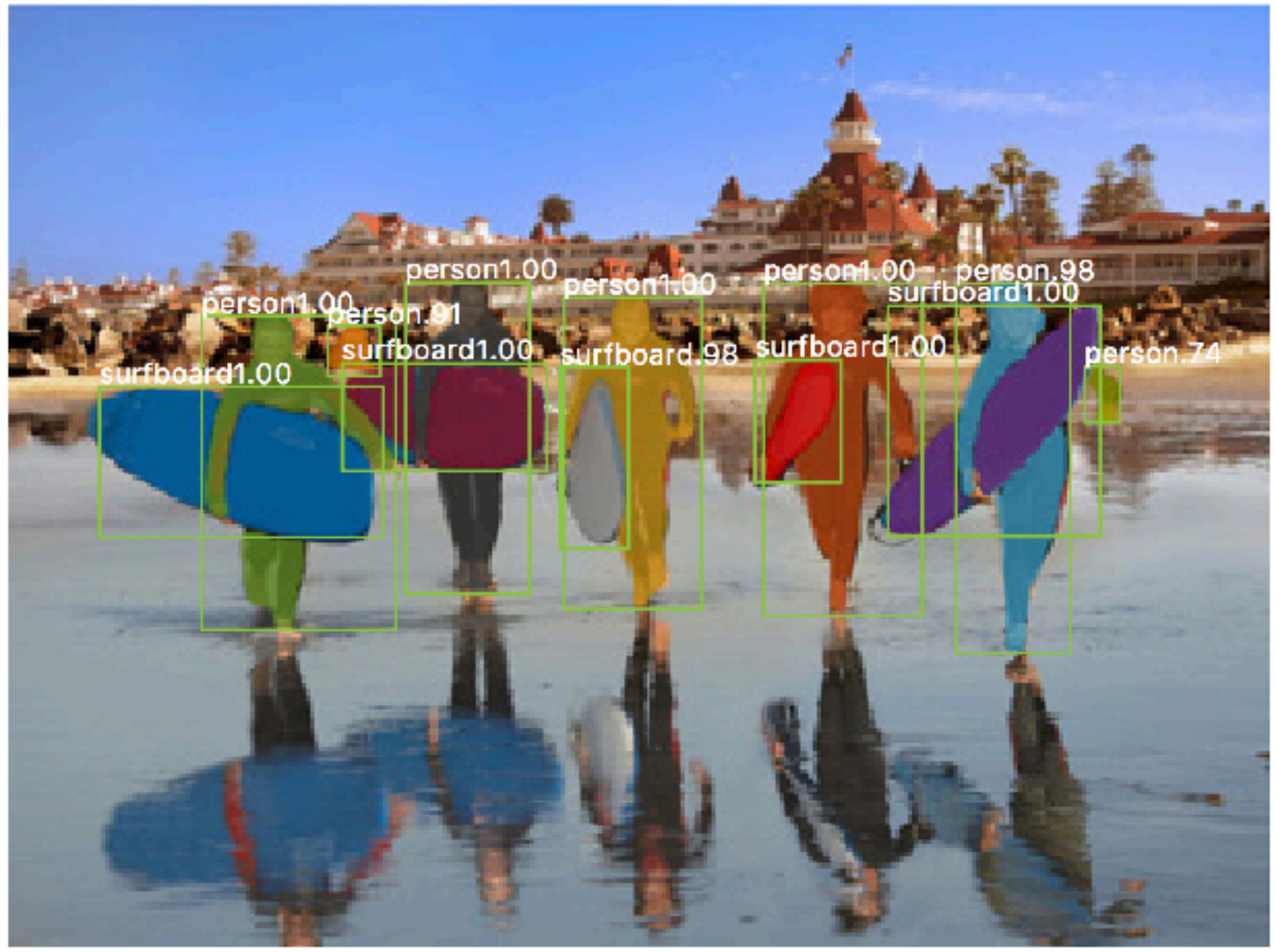
Date Issued: 1966-07-01

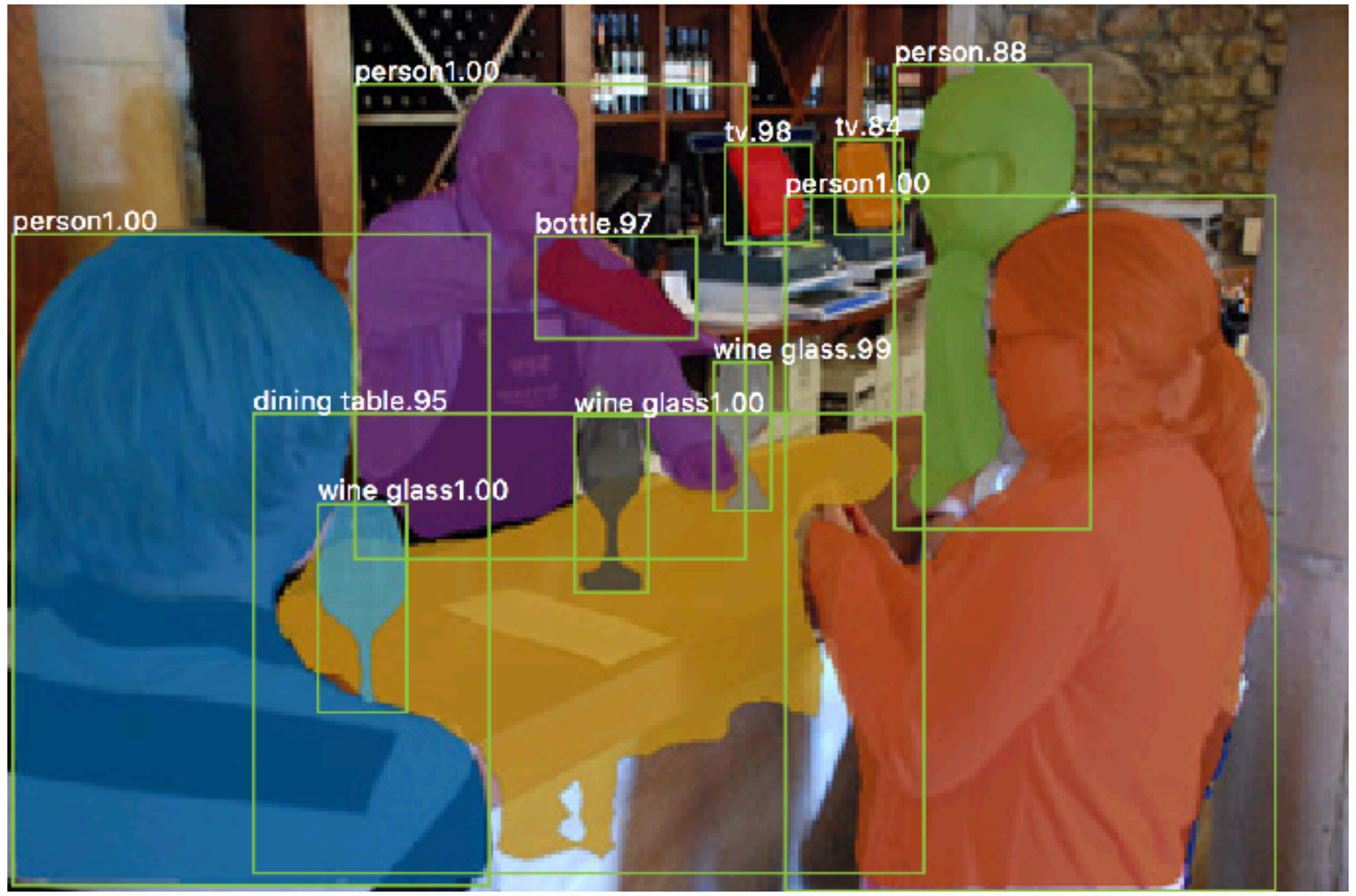
Abstract:

The summer vision project is an attempt to use our summer workers effectively in the construction of a significant part of a visual system. The particular task was chosen partly because it can be segmented into sub-problems which allow individuals to work independently and yet participate in the construction of a system complex enough to be real landmark in the development of "pattern recognition". The basic structure is fixed for the first phase of work extending to some point in July. Everyone is invited to contribute to the discussion of the second phase. Sussman is coordinator of "Vision Project" meetings and should be consulted by anyone who wishes to participate. The primary goal of the project is to construct a system of programs which will divide a vidisector picture into regions such as likely objects, likely background areas and chaos. We shall call this part of its operation FIGURE-GROUND analysis. It will be impossible to do this without considerable analysis of shape and surface properties, so FIGURE-GROUND analysis is really inseparable in practice from the second goal which is REGION DESCRIPTION. The final goal is OBJECT IDENTIFICATION which will actually name objects by matching them with a vocabulary of known objects.

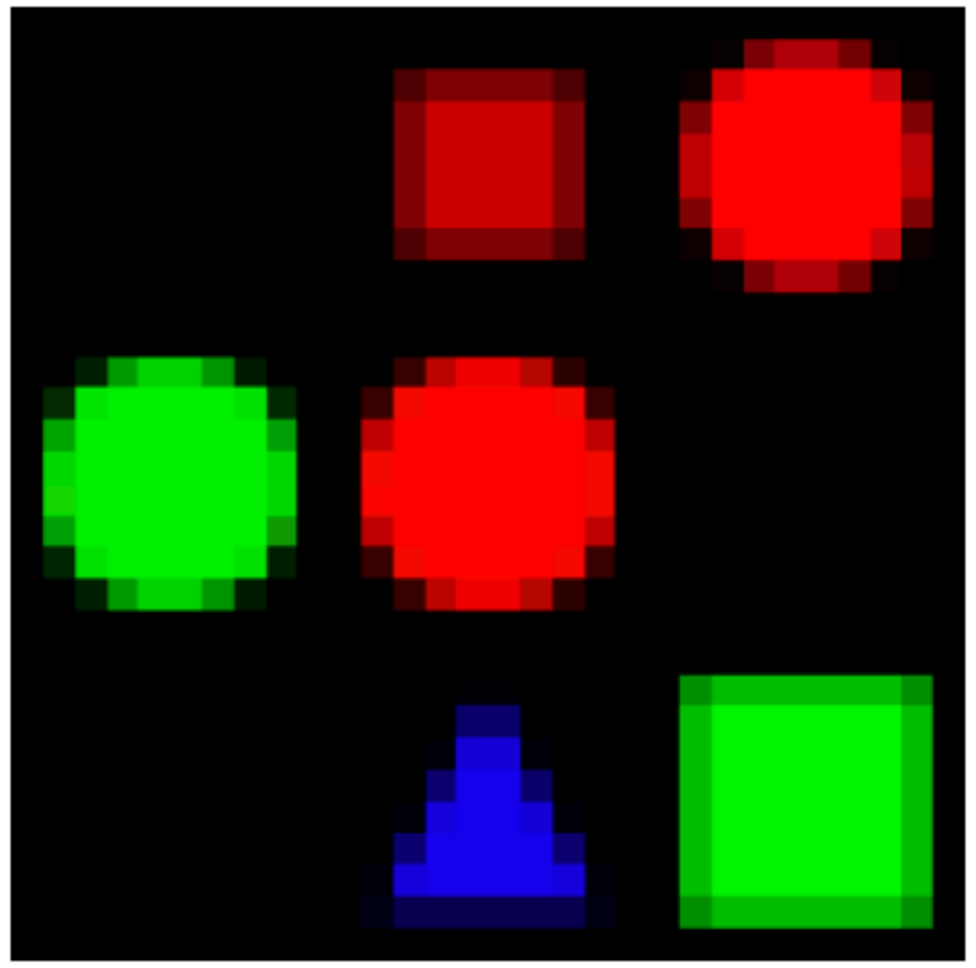
Just a few years ago…







[“Mask RCNN”, He et al. 2017]

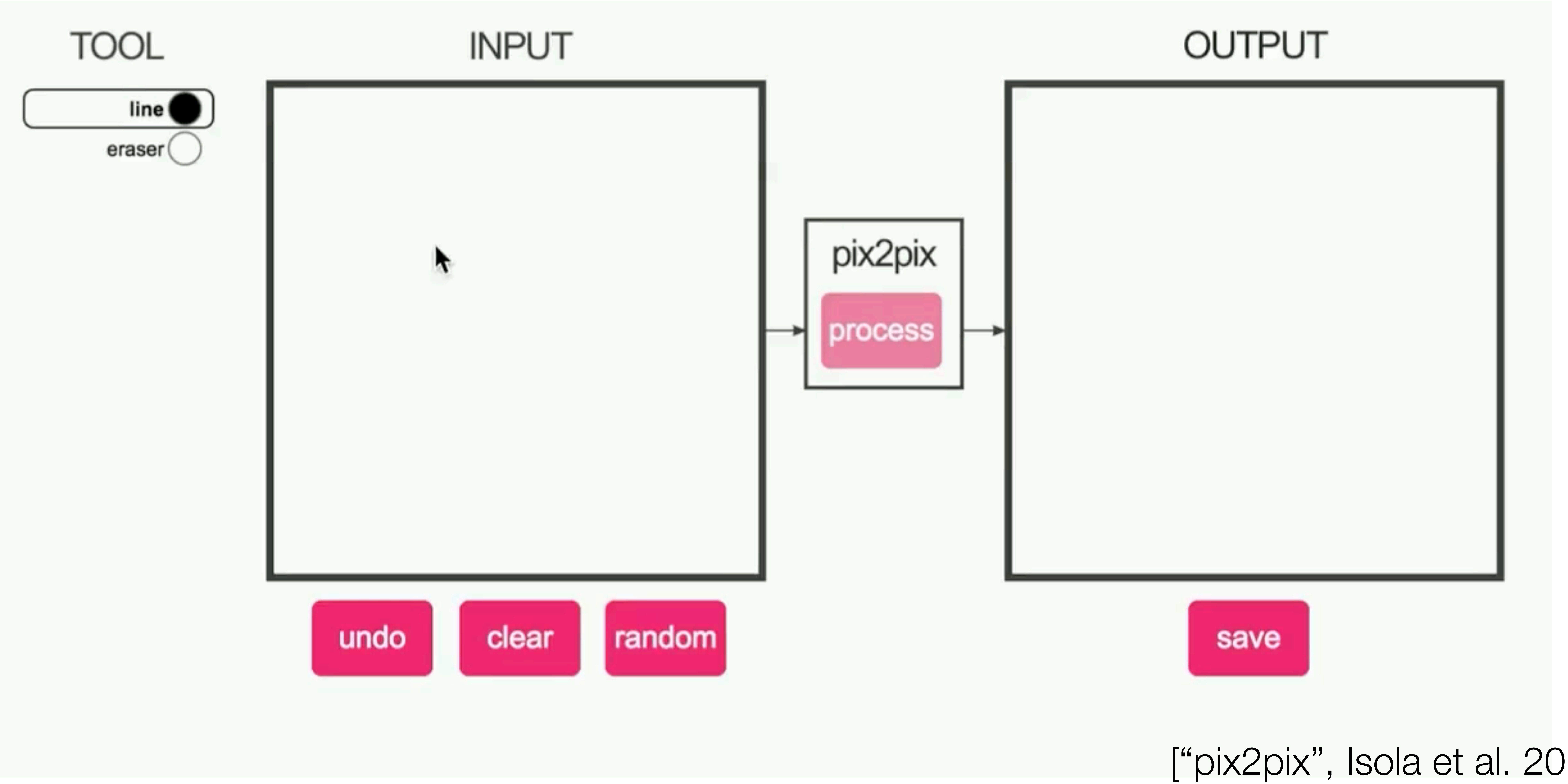
 *what color is the vase?*	 *is the bus full of passengers?*	 *is there a red shape above a circle?*

`classify[color](`
`  attend[vase])`
	`measure[is](`
	`  combine[and](`
	`    attend[bus],`
	`    attend[full])`
		`measure[is](`
		`  combine[and](`
		`   attend[red],`
		`   re-attend[above](`
		`    attend[circle])))`

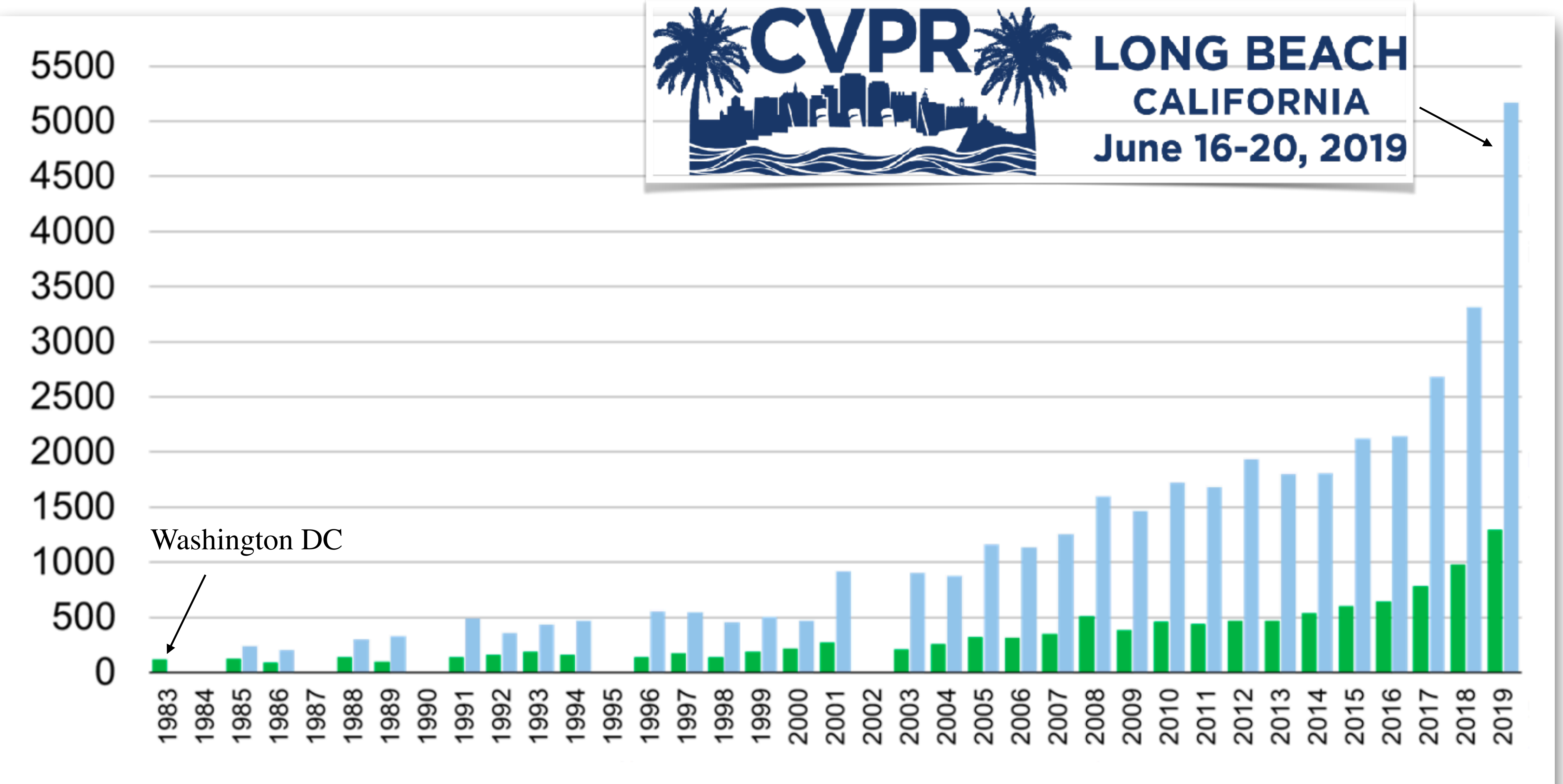
green (green)	yes (yes)	no (no)

[“Neural module networks”, Andreas et al. 2017]

#edges2cats [Chris Hesse]



[“pix2pix”, Isola et al. 2017]



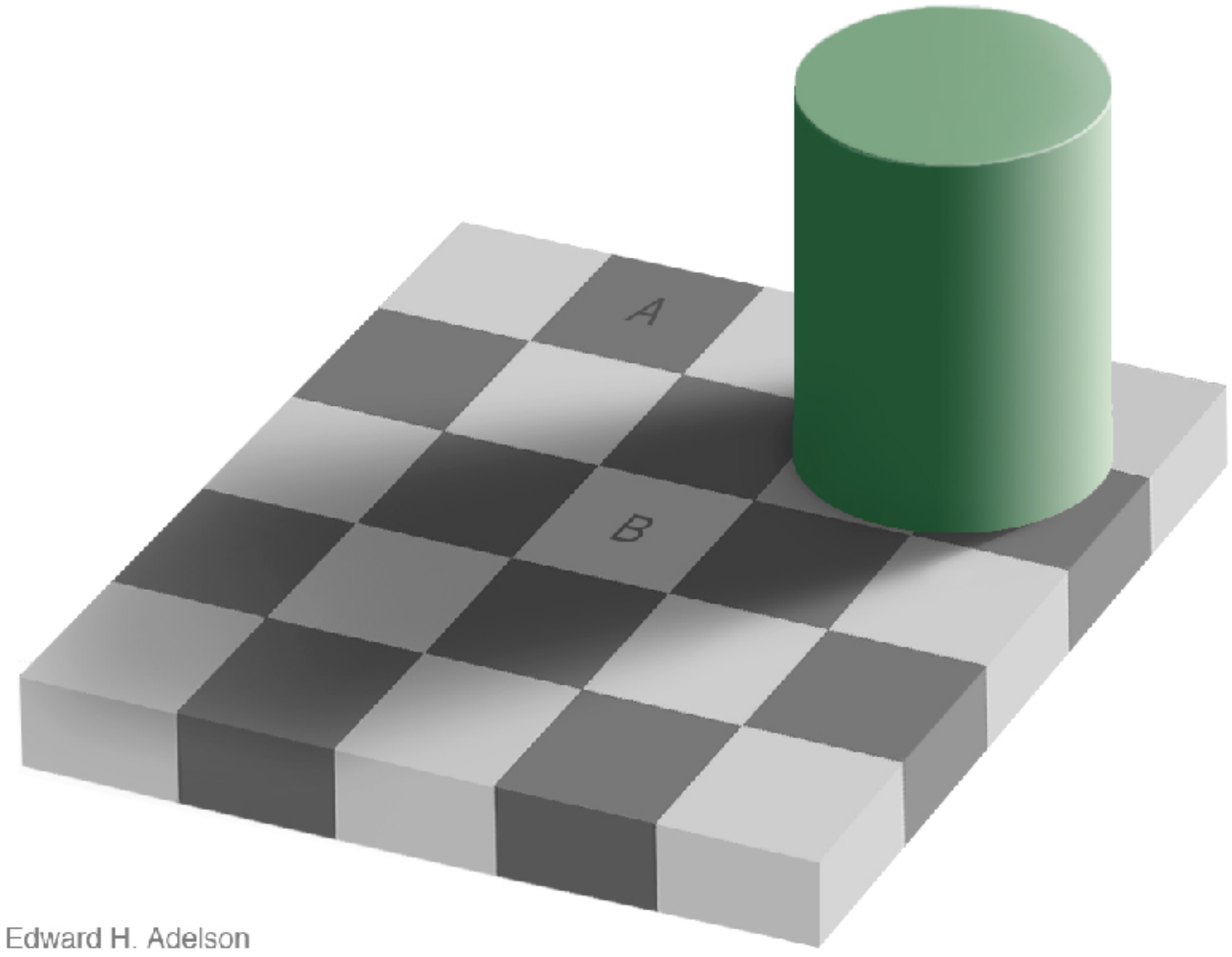
Number of submitted (blue) and accepted (green) papers in CVPR by year.

Source: CVPR 2019, Derek Hoiem

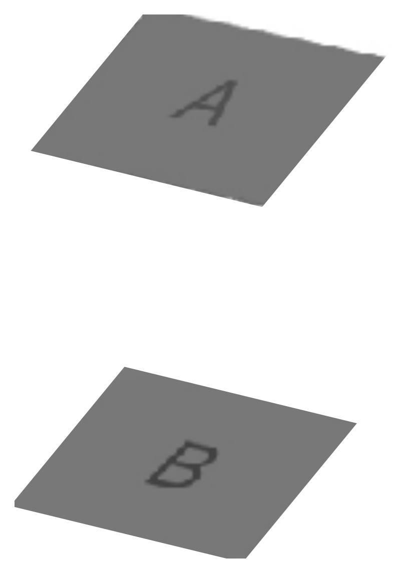
https://medium.com/reconstruct-inc/the-golden-age-of-computer-vision-338da3e471d1

Why is vision hard?

To see: perception vs. measurement



To see: perception vs. measurement



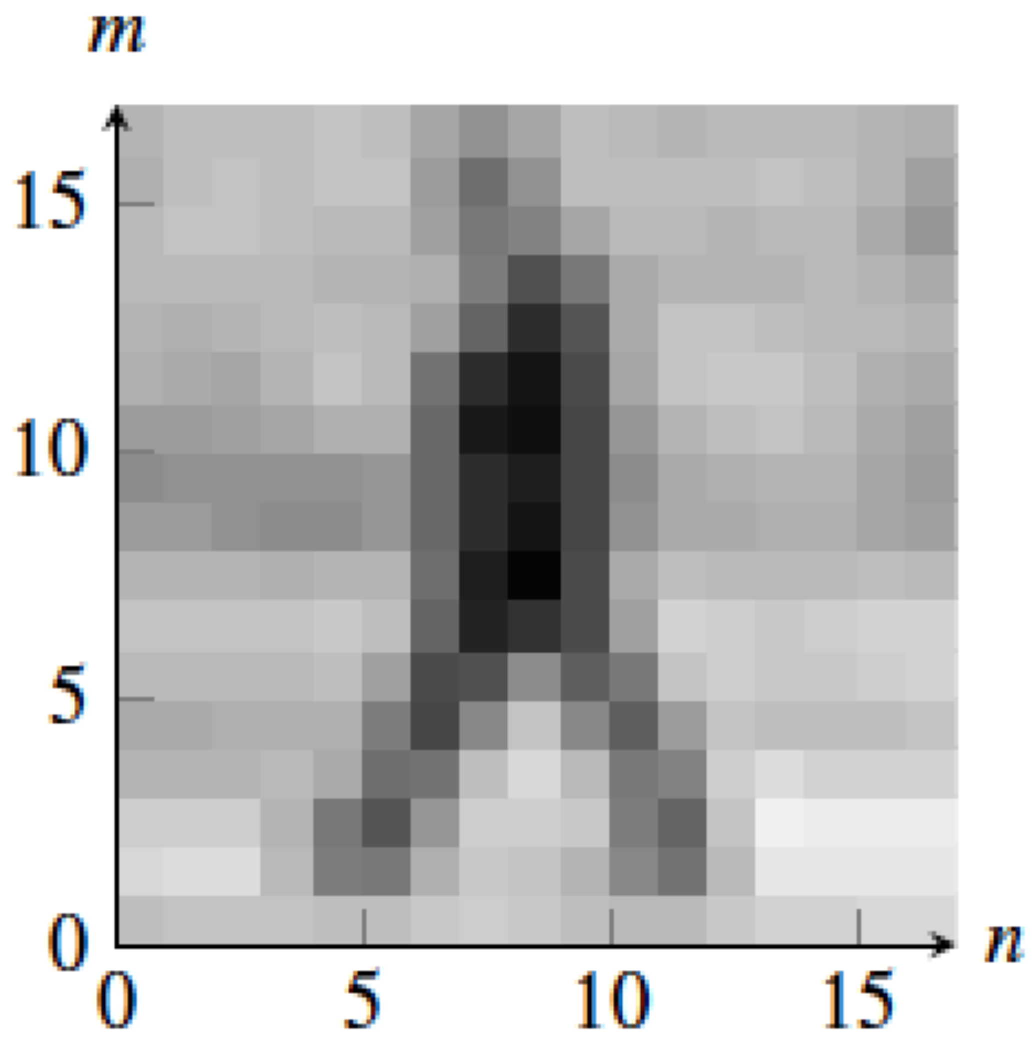
To see: perception vs. measurement

What the machine gets

$$
\mathbf{I} = \begin{bmatrix}
160 & 175 & 171 & 168 & 168 & 172 & 164 & 158 & 167 & 173 & 167 & 163 & 162 & 164 & 160 & 159 & 163 & 162 \\
149 & 164 & 172 & 175 & 178 & 179 & 176 & 118 & 97 & 168 & 175 & 171 & 169 & 175 & 176 & 177 & 165 & 152 \\
161 & 166 & 182 & 171 & 170 & 177 & 175 & 116 & 109 & 169 & 177 & 173 & 168 & 175 & 175 & 159 & 153 & 123 \\
171 & 174 & 177 & 175 & 167 & 161 & 157 & 138 & 103 & 112 & 157 & 164 & 159 & 160 & 165 & 169 & 148 & 144 \\
163 & 163 & 162 & 165 & 167 & 164 & 178 & 167 & 77 & 55 & 134 & 170 & 167 & 162 & 164 & 175 & 168 & 160 \\
173 & 164 & 158 & 165 & 180 & 180 & 150 & 89 & 61 & 34 & 137 & 186 & 186 & 182 & 175 & 165 & 160 & 164 \\
152 & 155 & 146 & 147 & 169 & 180 & 163 & 51 & 24 & 32 & 119 & 163 & 175 & 182 & 181 & 162 & 148 & 153 \\
134 & 135 & 147 & 149 & 150 & 147 & 148 & 62 & 36 & 46 & 114 & 157 & 163 & 167 & 169 & 163 & 146 & 147 \\
135 & 132 & 131 & 125 & 115 & 129 & 132 & 74 & 54 & 41 & 104 & 156 & 152 & 156 & 164 & 156 & 141 & 144 \\
151 & 155 & 151 & 145 & 144 & 149 & 143 & 71 & 31 & 29 & 129 & 164 & 157 & 155 & 159 & 158 & 156 & 148 \\
172 & 174 & 178 & 177 & 177 & 181 & 174 & 54 & 21 & 29 & 136 & 190 & 180 & 179 & 176 & 184 & 187 & 182 \\
177 & 178 & 176 & 173 & 174 & 180 & 150 & 27 & 101 & 94 & 74 & 189 & 188 & 186 & 183 & 186 & 188 & 187 \\
160 & 160 & 163 & 163 & 161 & 167 & 100 & 45 & 169 & 166 & 59 & 136 & 184 & 176 & 175 & 177 & 185 & 186 \\
147 & 150 & 153 & 155 & 160 & 155 & 56 & 111 & 182 & 180 & 104 & 84 & 168 & 172 & 171 & 164 & 168 & 167 \\
184 & 182 & 178 & 175 & 179 & 133 & 86 & 191 & 201 & 204 & 191 & 79 & 172 & 220 & 217 & 205 & 209 & 200 \\
184 & 187 & 192 & 182 & 124 & 32 & 109 & 168 & 171 & 167 & 163 & 51 & 105 & 203 & 209 & 203 & 210 & 205 \\
191 & 198 & 203 & 197 & 175 & 149 & 169 & 189 & 190 & 173 & 160 & 145 & 156 & 202 & 199 & 201 & 205 & 202 \\
153 & 149 & 153 & 155 & 173 & 182 & 179 & 177 & 182 & 177 & 182 & 185 & 179 & 177 & 167 & 176 & 182 & 180
\end{bmatrix}
$$

The camera is a measurement device, not a vision system

What we see



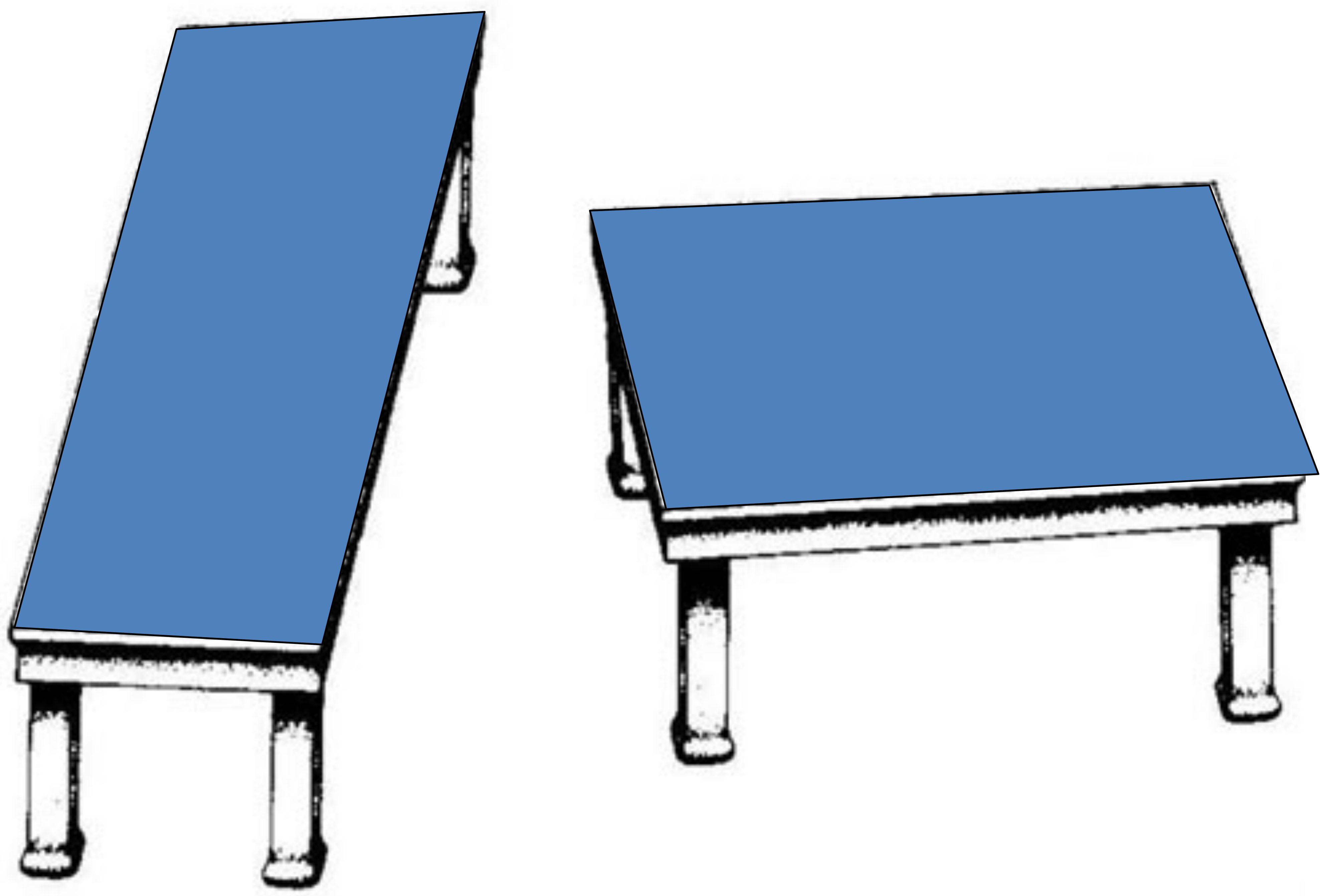
What the machine gets

$$
\mathbf{I} = \begin{bmatrix}
160 & 175 & 171 & 168 & 168 & 172 & 164 & 158 & 167 & 173 & 167 & 163 & 162 & 164 & 160 & 159 & 163 & 162 \\
149 & 164 & 172 & 175 & 178 & 179 & 176 & 118 & 97 & 168 & 175 & 171 & 169 & 175 & 176 & 177 & 165 & 152 \\
161 & 166 & 182 & 171 & 170 & 177 & 175 & 116 & 109 & 169 & 177 & 173 & 168 & 175 & 175 & 159 & 153 & 123 \\
171 & 174 & 177 & 175 & 167 & 161 & 157 & 138 & 103 & 112 & 157 & 164 & 159 & 160 & 165 & 169 & 148 & 144 \\
163 & 163 & 162 & 165 & 167 & 164 & 178 & 167 & 77 & 55 & 134 & 170 & 167 & 162 & 164 & 175 & 168 & 160 \\
173 & 164 & 158 & 165 & 180 & 180 & 150 & 89 & 61 & 34 & 137 & 186 & 186 & 182 & 175 & 165 & 160 & 164 \\
152 & 155 & 146 & 147 & 169 & 180 & 163 & 51 & 24 & 32 & 119 & 163 & 175 & 182 & 181 & 162 & 148 & 153 \\
134 & 135 & 147 & 149 & 150 & 147 & 148 & 62 & 36 & 46 & 114 & 157 & 163 & 167 & 169 & 163 & 146 & 147 \\
135 & 132 & 131 & 125 & 115 & 129 & 132 & 74 & 54 & 41 & 104 & 156 & 152 & 156 & 164 & 156 & 141 & 144 \\
151 & 155 & 151 & 145 & 144 & 149 & 143 & 71 & 31 & 29 & 129 & 164 & 157 & 155 & 159 & 158 & 156 & 148 \\
172 & 174 & 178 & 177 & 177 & 181 & 174 & 54 & 21 & 29 & 136 & 190 & 180 & 179 & 176 & 184 & 187 & 182 \\
177 & 178 & 176 & 173 & 174 & 180 & 150 & 27 & 101 & 94 & 74 & 189 & 188 & 186 & 183 & 186 & 188 & 187 \\
160 & 160 & 163 & 163 & 161 & 167 & 100 & 45 & 169 & 166 & 59 & 136 & 184 & 176 & 175 & 177 & 185 & 186 \\
147 & 150 & 153 & 155 & 160 & 155 & 56 & 111 & 182 & 180 & 104 & 84 & 168 & 172 & 171 & 164 & 168 & 167 \\
184 & 182 & 178 & 175 & 179 & 133 & 86 & 191 & 201 & 204 & 191 & 79 & 172 & 220 & 217 & 205 & 209 & 200 \\
184 & 187 & 192 & 182 & 124 & 32 & 109 & 168 & 171 & 167 & 163 & 51 & 105 & 203 & 209 & 203 & 210 & 205 \\
191 & 198 & 203 & 197 & 175 & 149 & 169 & 189 & 190 & 173 & 160 & 145 & 156 & 202 & 199 & 201 & 205 & 202 \\
153 & 149 & 153 & 155 & 173 & 182 & 179 & 177 & 182 & 177 & 182 & 185 & 179 & 177 & 167 & 176 & 182 & 180
\end{bmatrix}
$$

The camera is a measurement device, not a vision system

To see: perception vs. measurement

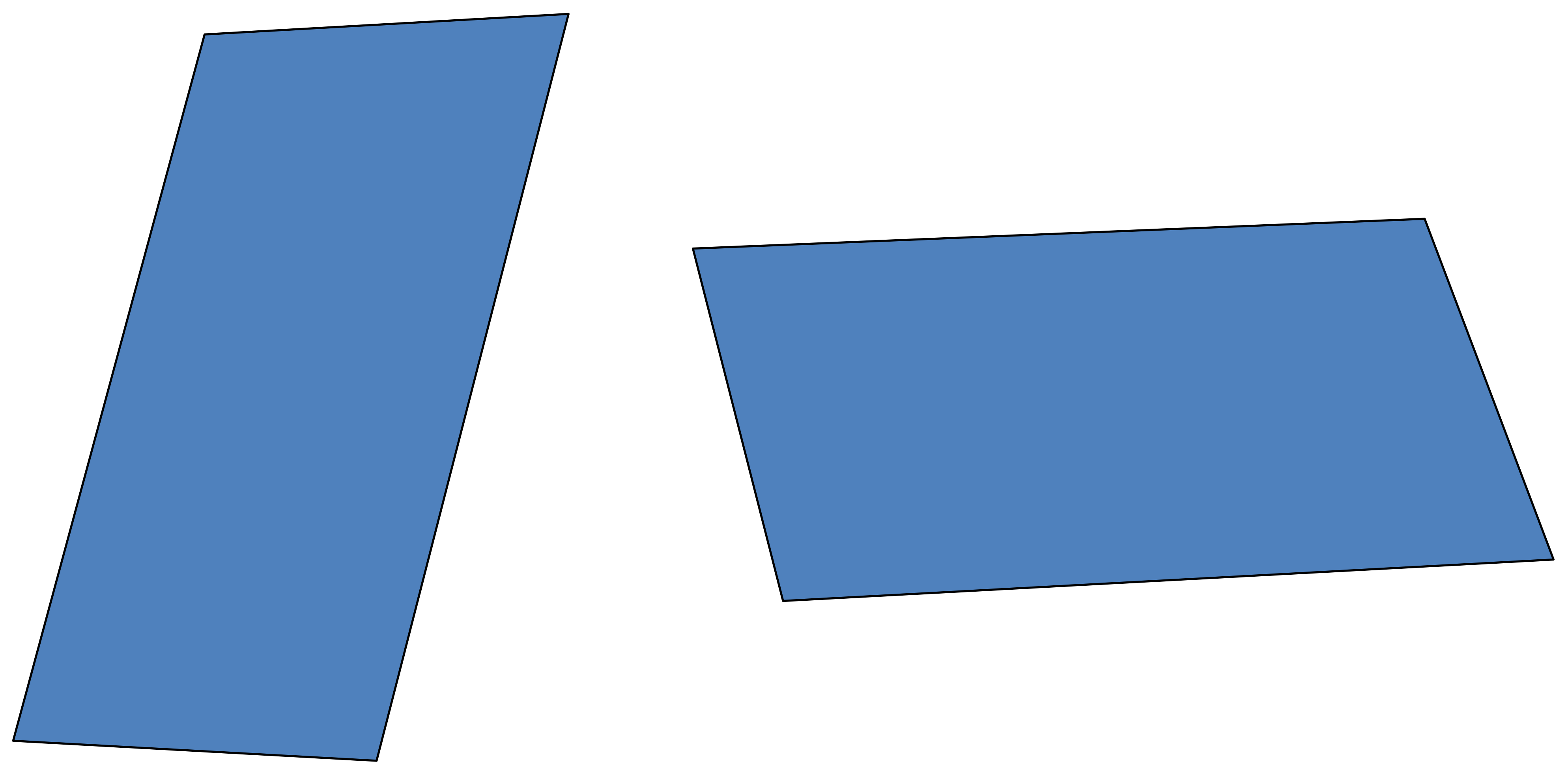
Depth processing is automatic, and we can not shut it down…



by Roger Shepard ("Turning the Tables")

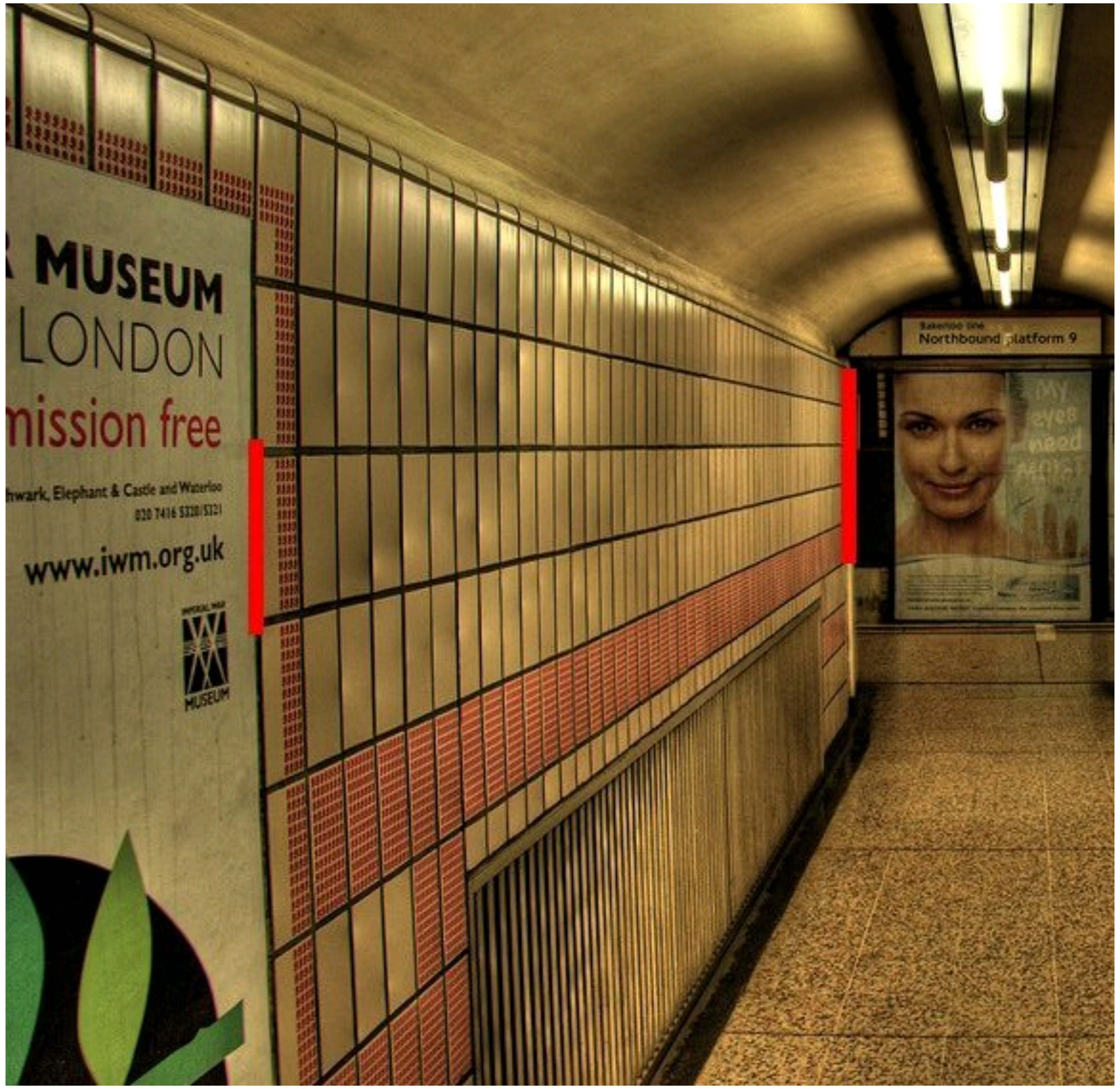
To see: perception vs. measurement

Depth processing is automatic, and we can not shut it down…



by Roger Shepard (”Turning the Tables”)

To see: perception vs. measurement



(c) 2006 Walt Anthony

To see: perception vs. measurement

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

PROJECT MAC

Artificial Intelligence Group

Vision Memo. No. 100.

July 7, 1966

THE SUMMER VISION PROJECT

Seymour Papert

The summer vision project is an attempt to use our summer workers effectively in the construction of a significant part of a visual system. The particular task was chosen partly because it can be segmented into sub-problems which will allow individuals to work independently and yet participate in the construction of a system complex enough to be a real landmark in the development of "pattern recognition".

Problem set 1
The “one week” vision project

The goal of the first problem set is
to solve vision

A Simple Visual System

- A simple world
- A simple image formation model
- A simple goal

A Simple World

A Simple World

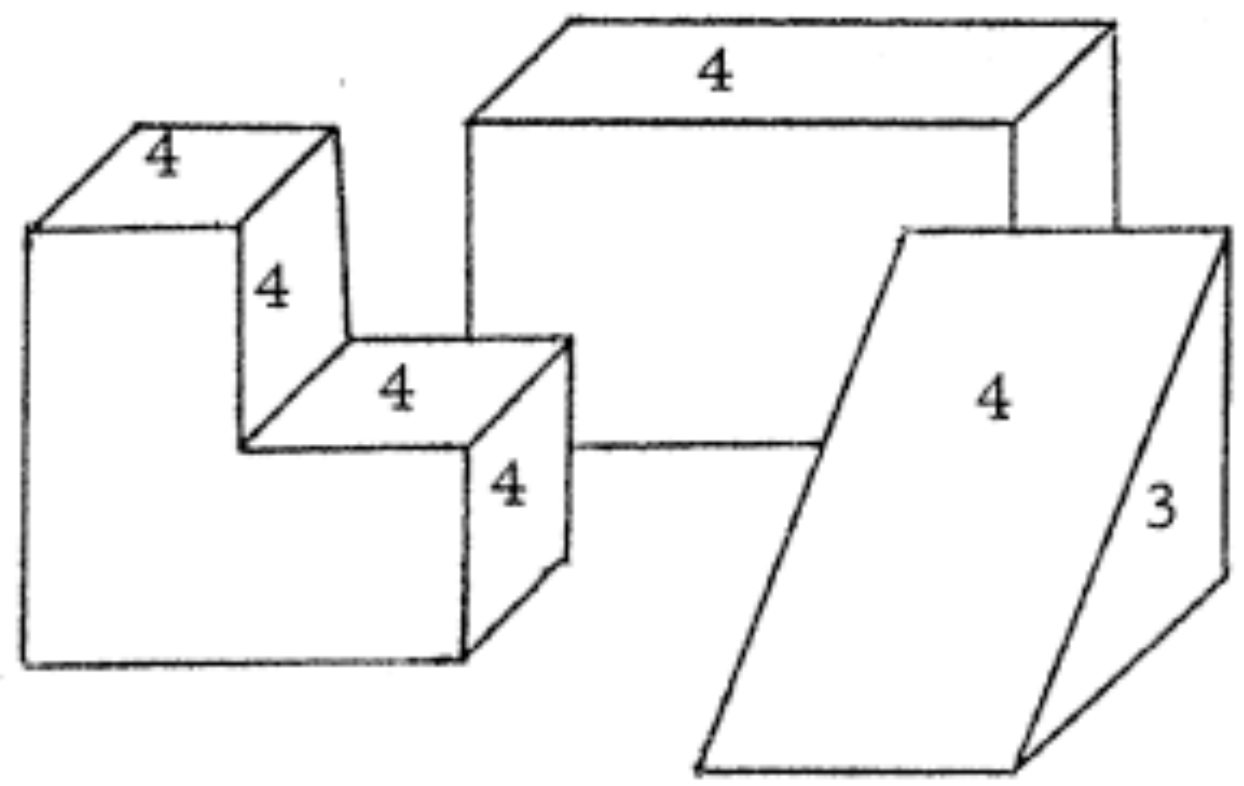
MACHINE PERCEPTION OF THREE-DIMENSIONAL SOLIDS

by

LAWRENCE GILMAN ROBERTS

Submitted to the Department of Electrical Engineering on May 10, 1963, in partial fulfillment of the requirements for the degree of Doctor of Philosophy.

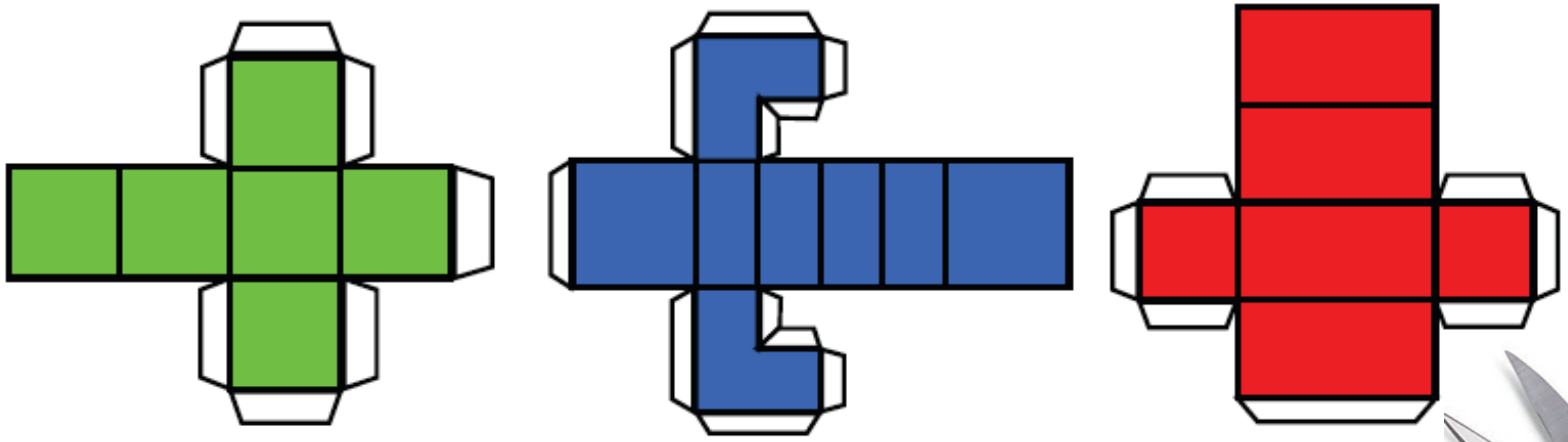
The problem of machine recognition of pictorial data has long been a challenging goal, but has seldom been attempted with anything more complex than alphabetic characters. Many people have felt that research on character recognition would be a first step, leading the way to a more general pattern recognition system. However, the multitudinous attempts at character recognition, including my own, have not led very far. The reason, I feel, is that the study of abstract, two-dimensional forms leads us away from, not toward, the techniques necessary for the recognition of three-dimensional objects. The per-

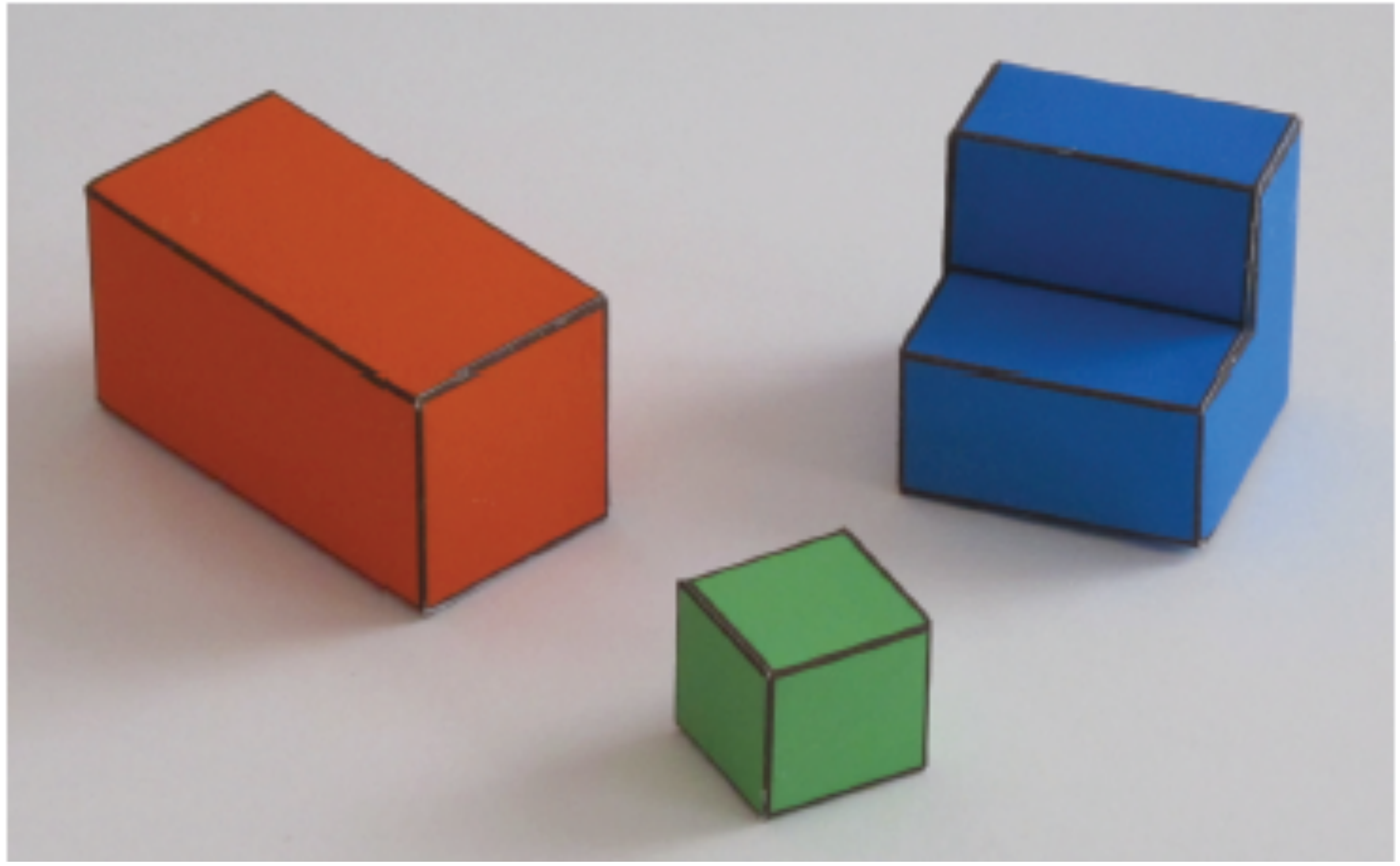


Complete Convex Polygons. The polygon selection procedure would select the numbered polygons as complete and convex. The number indicates the probable number of sides. A polygon is incomplete if one of its points is a collinear joint of another polygon.

http://www.packet.cc/files/mach-per-3D-solids.html

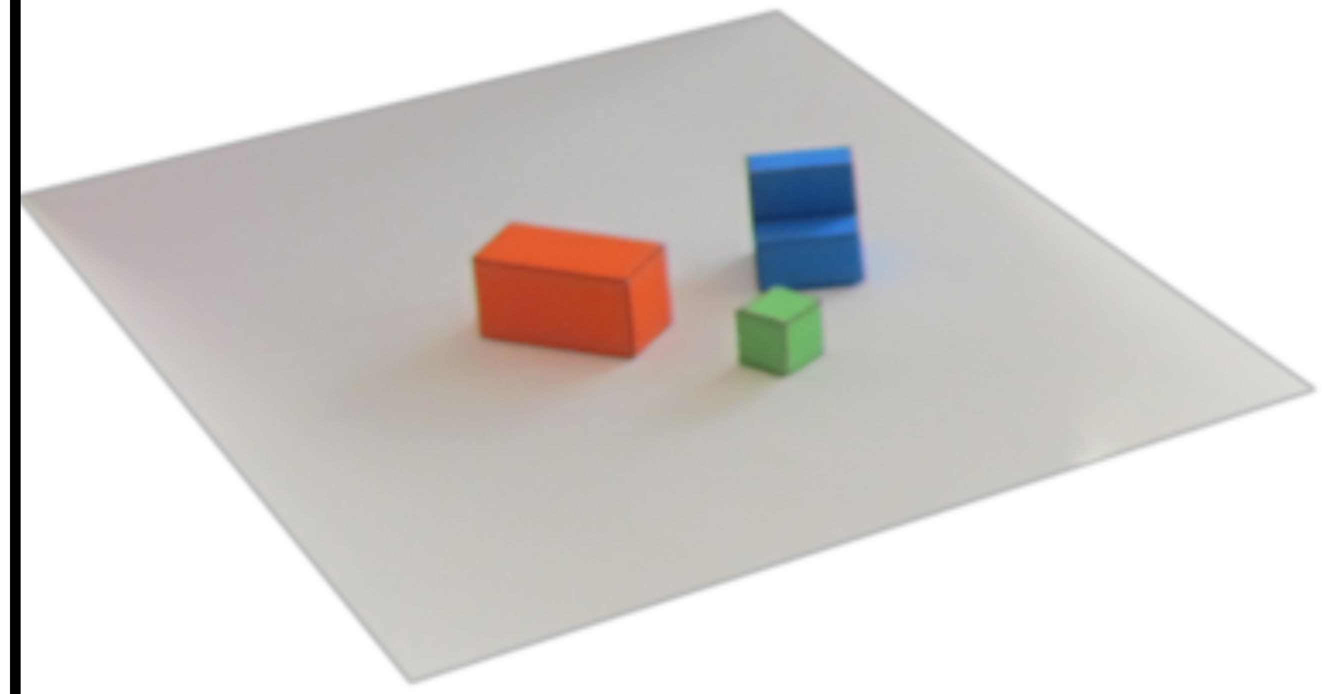
A Simple World



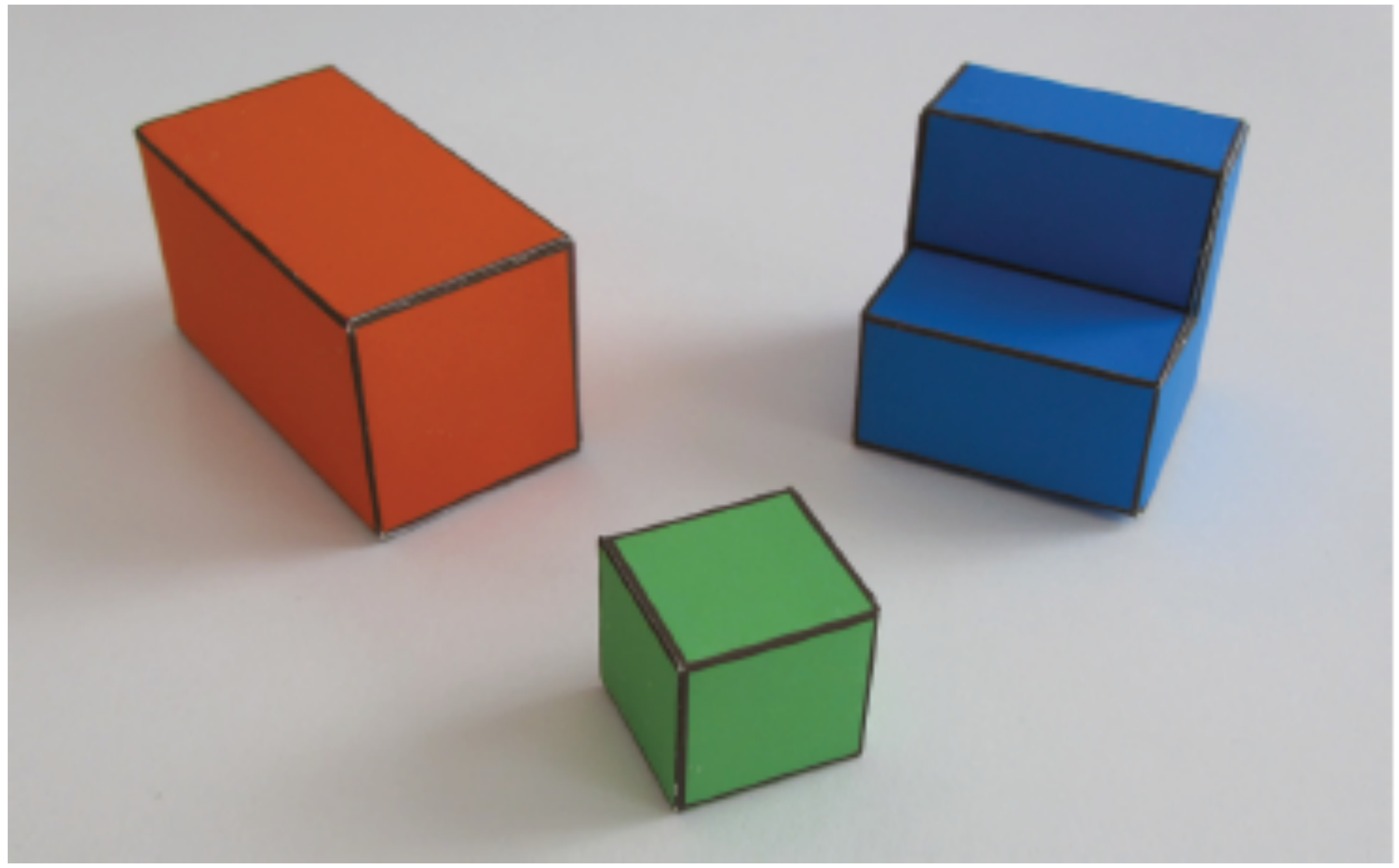


A simple image formation model

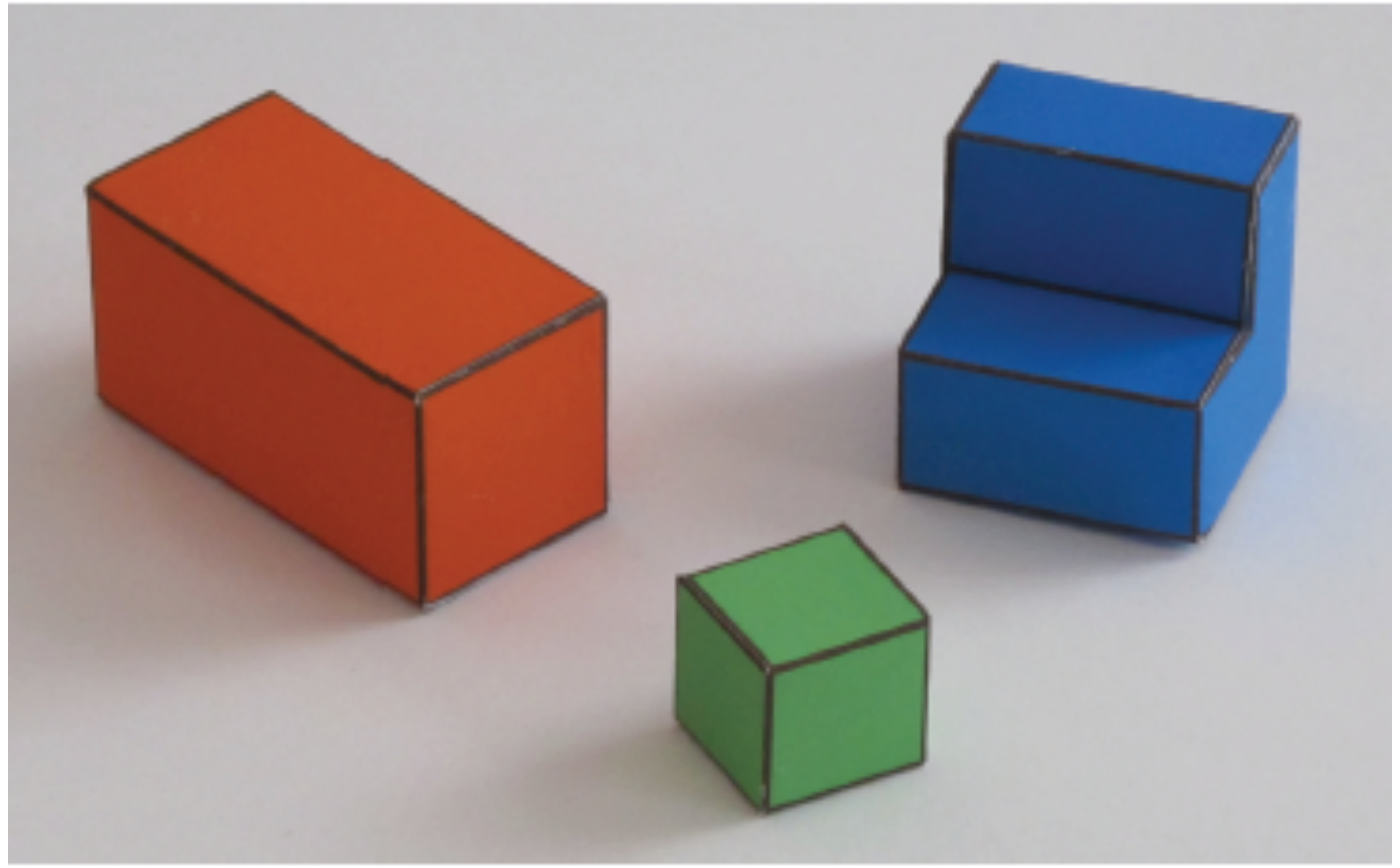
Simple world rules:
- Surfaces can be horizontal or vertical.
- Objects will be resting on a white horizontal ground plane



A simple image formation model

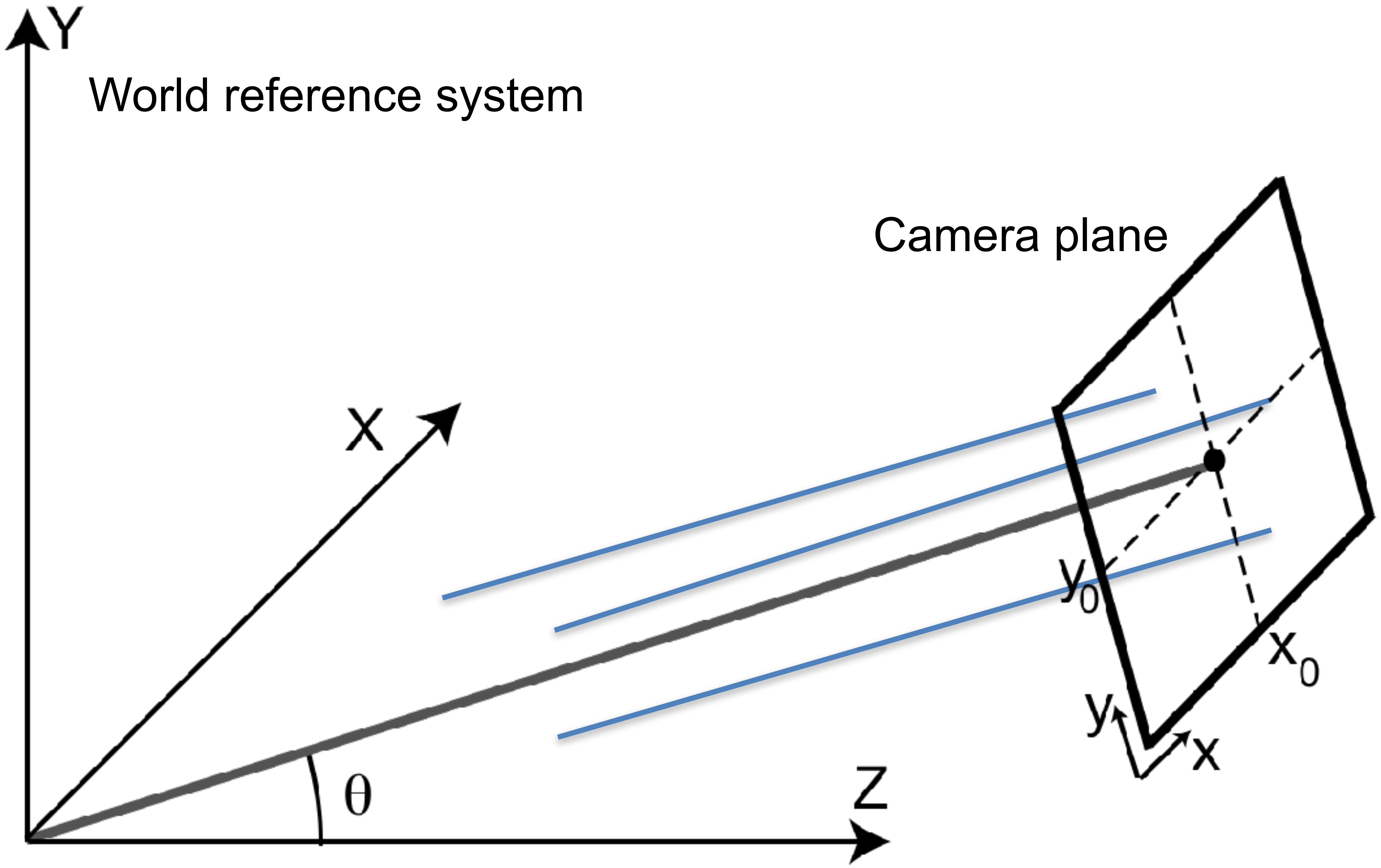


Perspective projection



Parallel (orthographic) projection

A simple image formation model



(right-handed reference system)

A simple image formation model

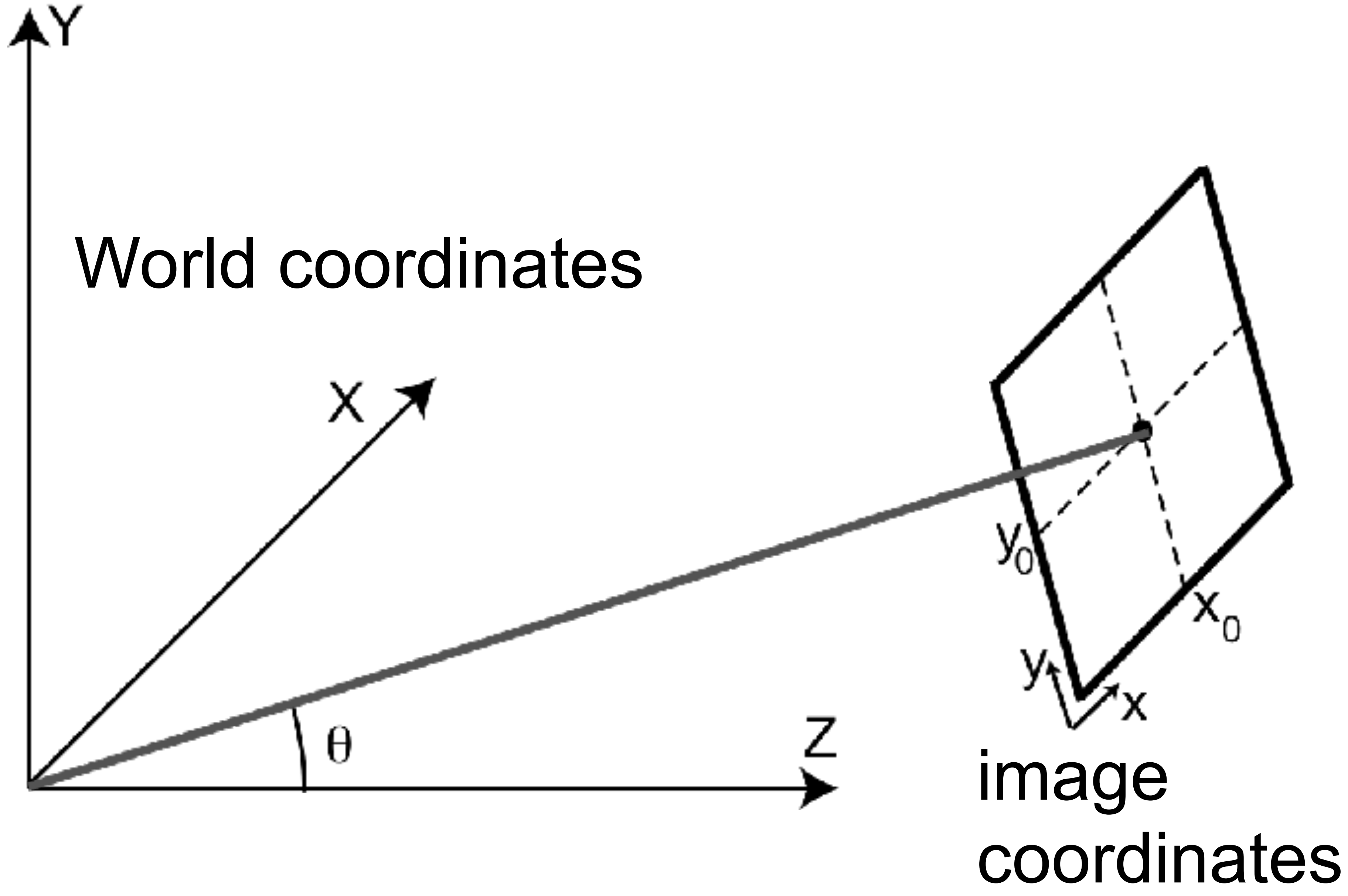
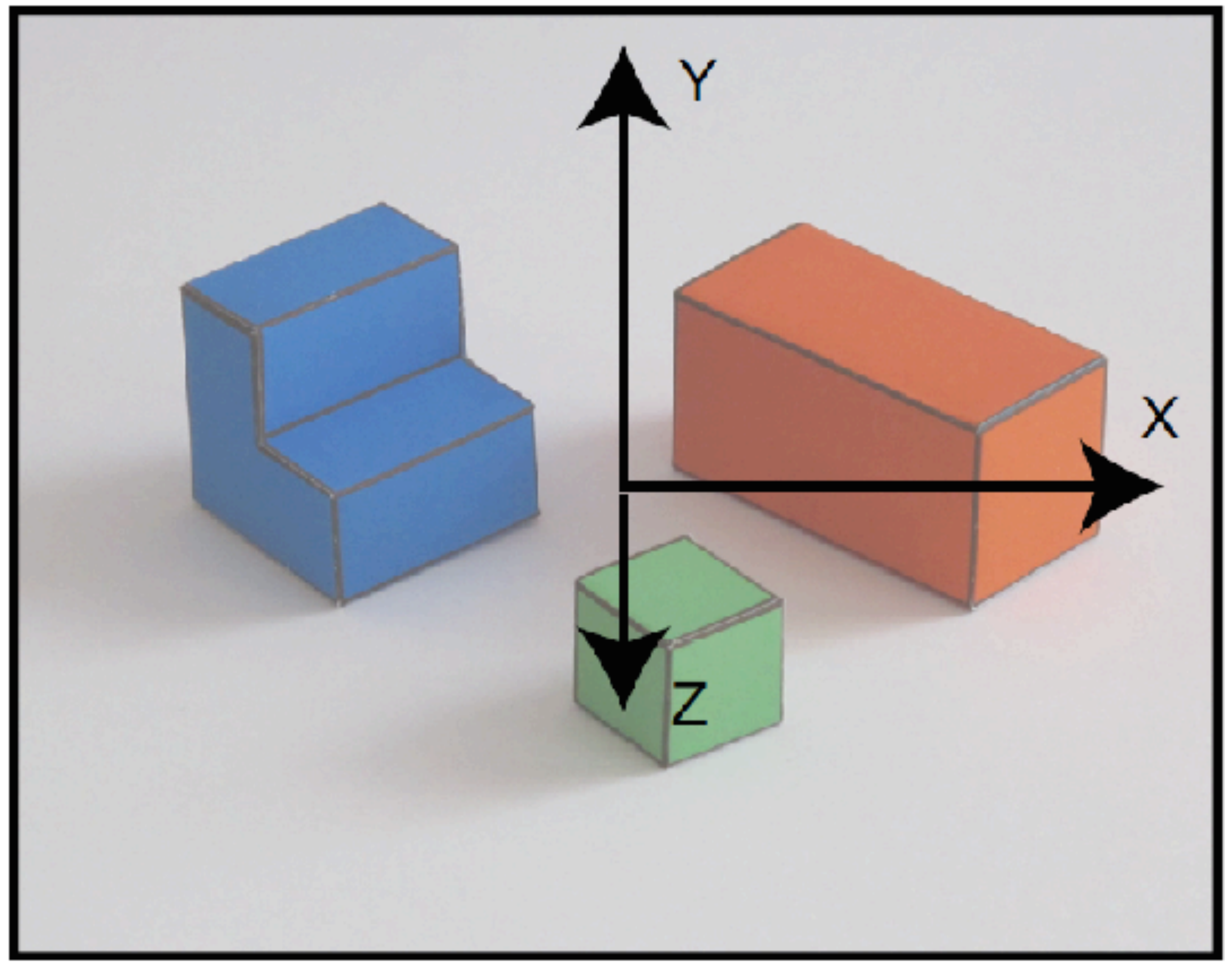


Image and projection of the world coordinate axes into the image plane



World coordinates

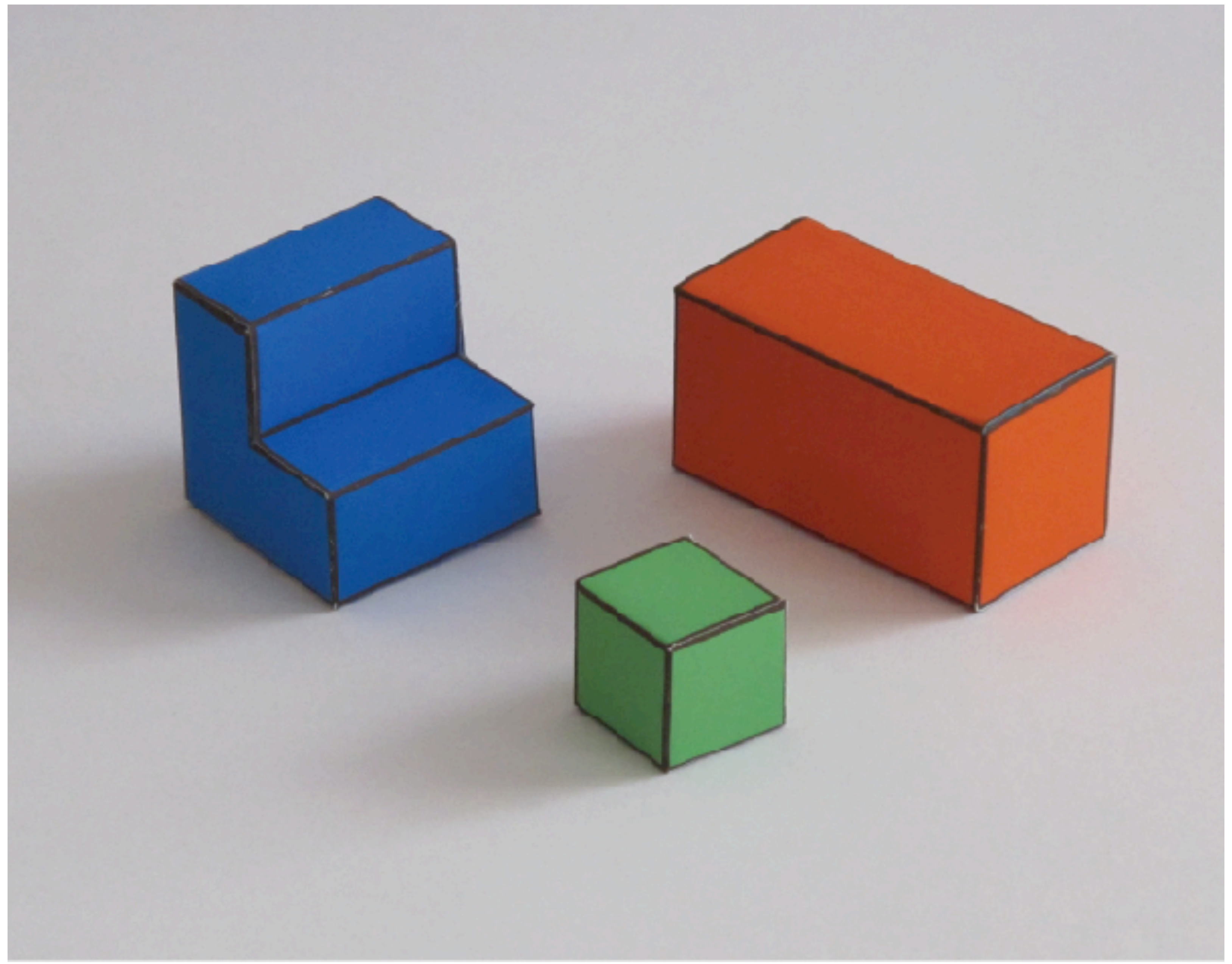
$$x = X + x_0$$

$$y = \cos(\theta)\, Y - \sin(\theta)\, Z + y_0$$

image coordinates

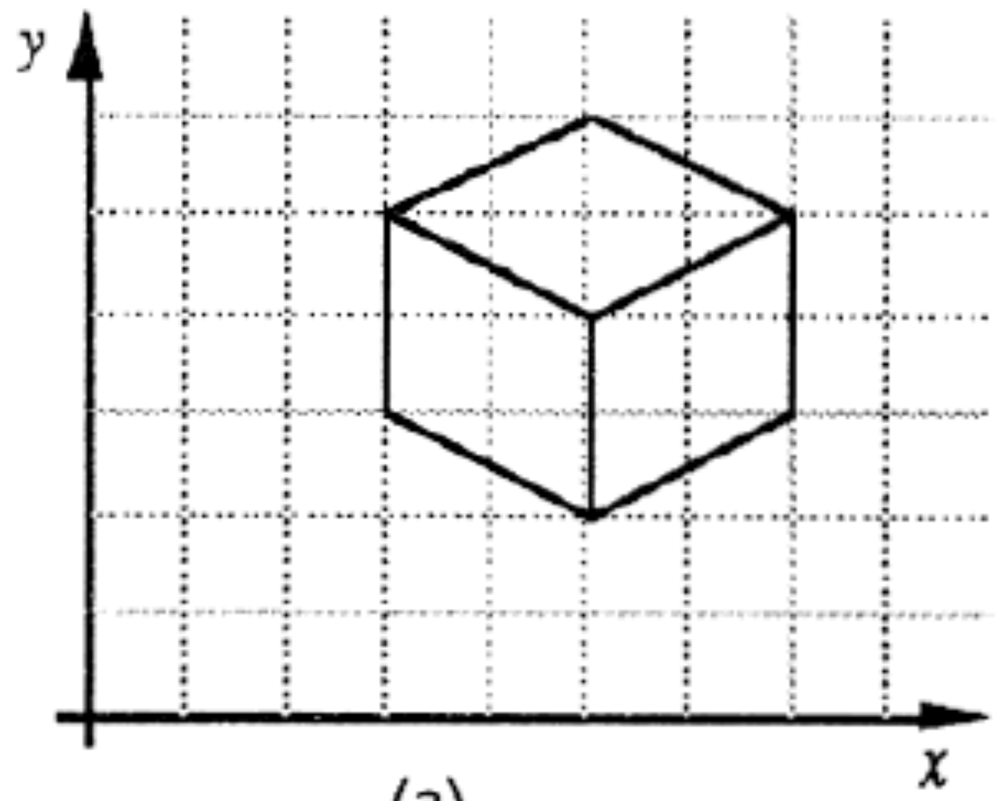
A simple goal

To recover the 3D structure of the world



We want to recover X(x,y), Y(x,y), Z(x,y) using as input I(x,y)

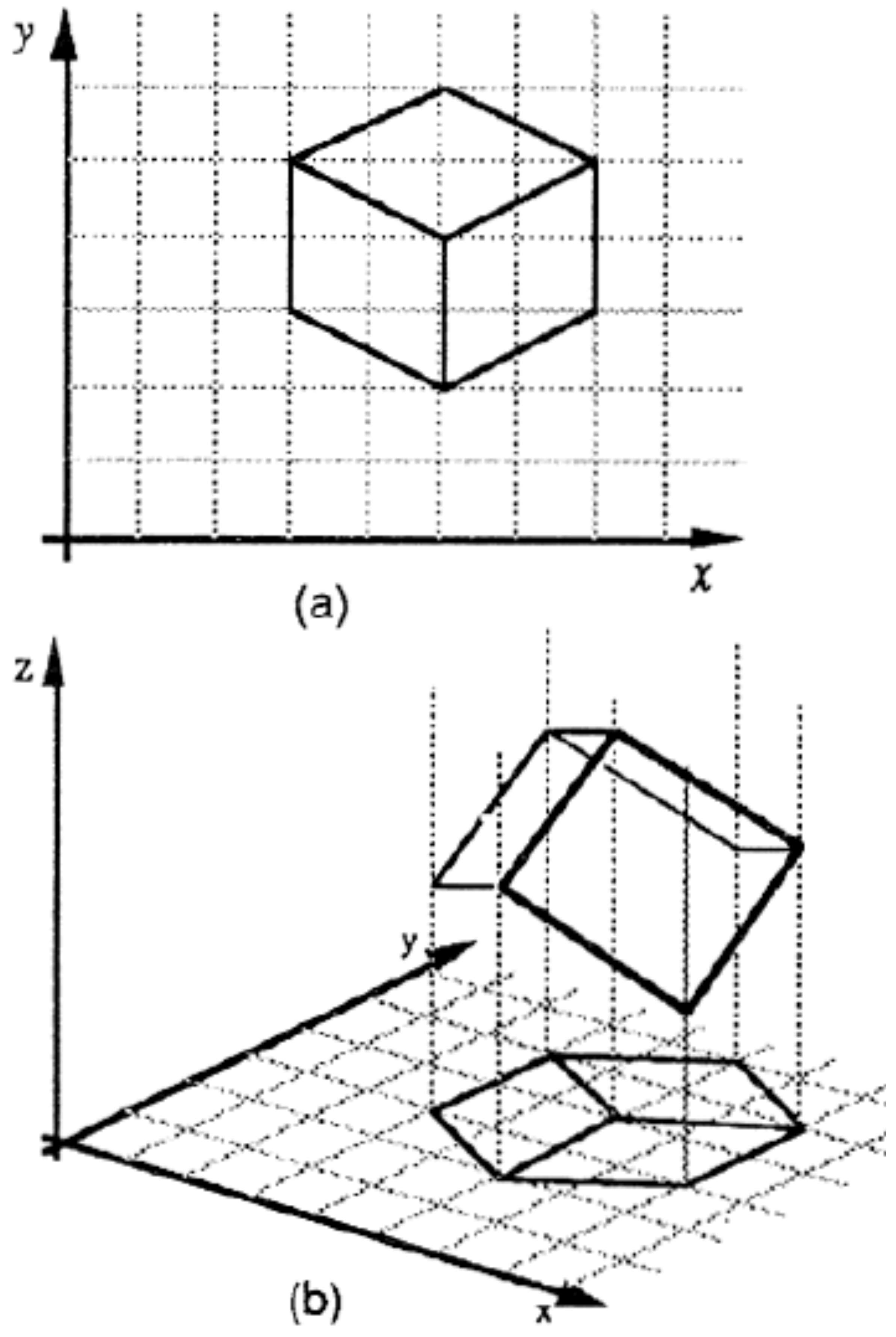
Why is this hard?



(a)

Sinha & Adelson 93

Why is this hard?



Sinha & Adelson 93

Why is this hard?

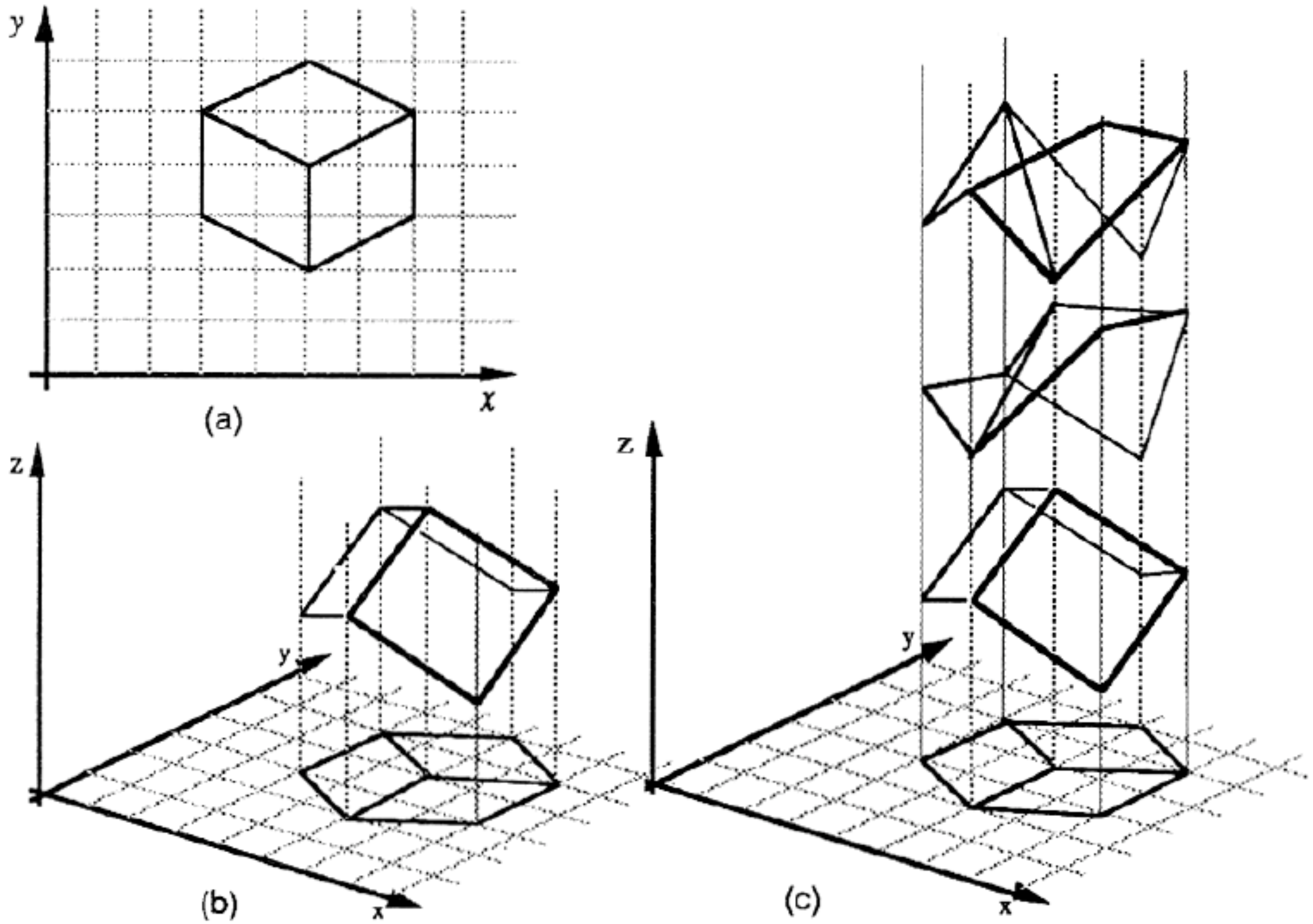
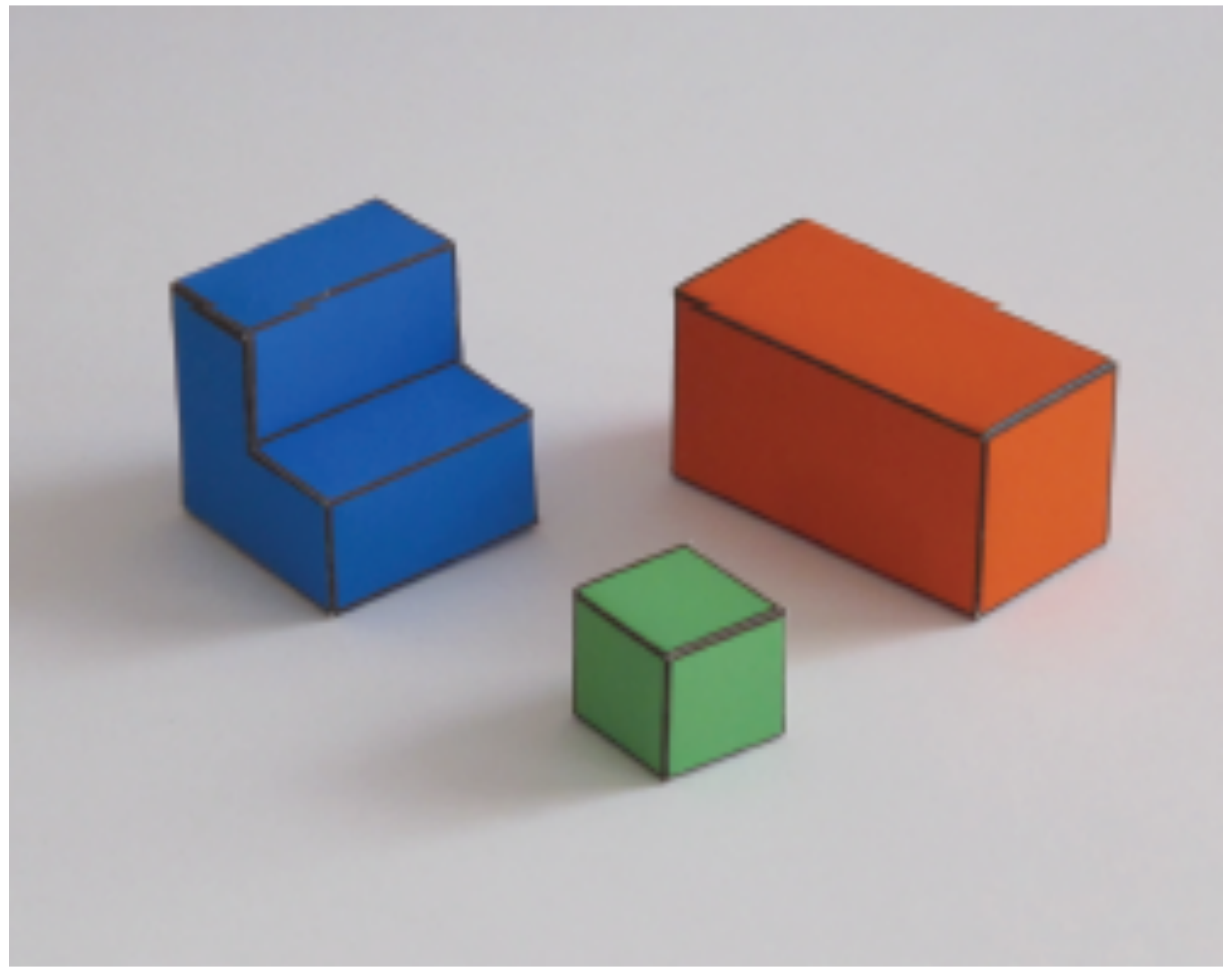
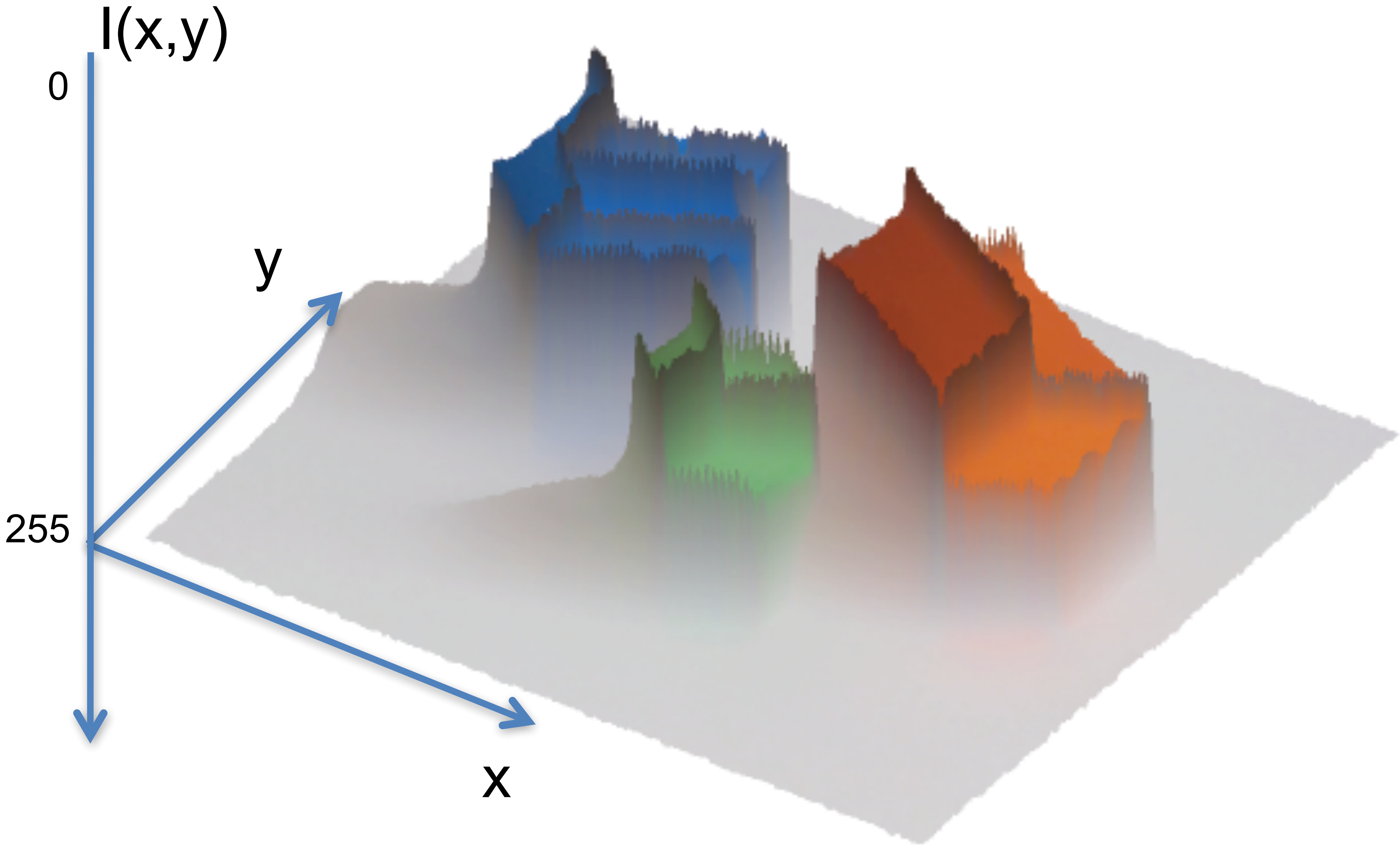


Figure 1. (a) A line drawing provides information only about the x, y coordinates of points lying along the object contours. (b) The human visual system is usually able to reconstruct an object in three dimensions given only a single 2D projection (c) Any planar line-drawing is geometrically consistent with infinitely many 3D structures.

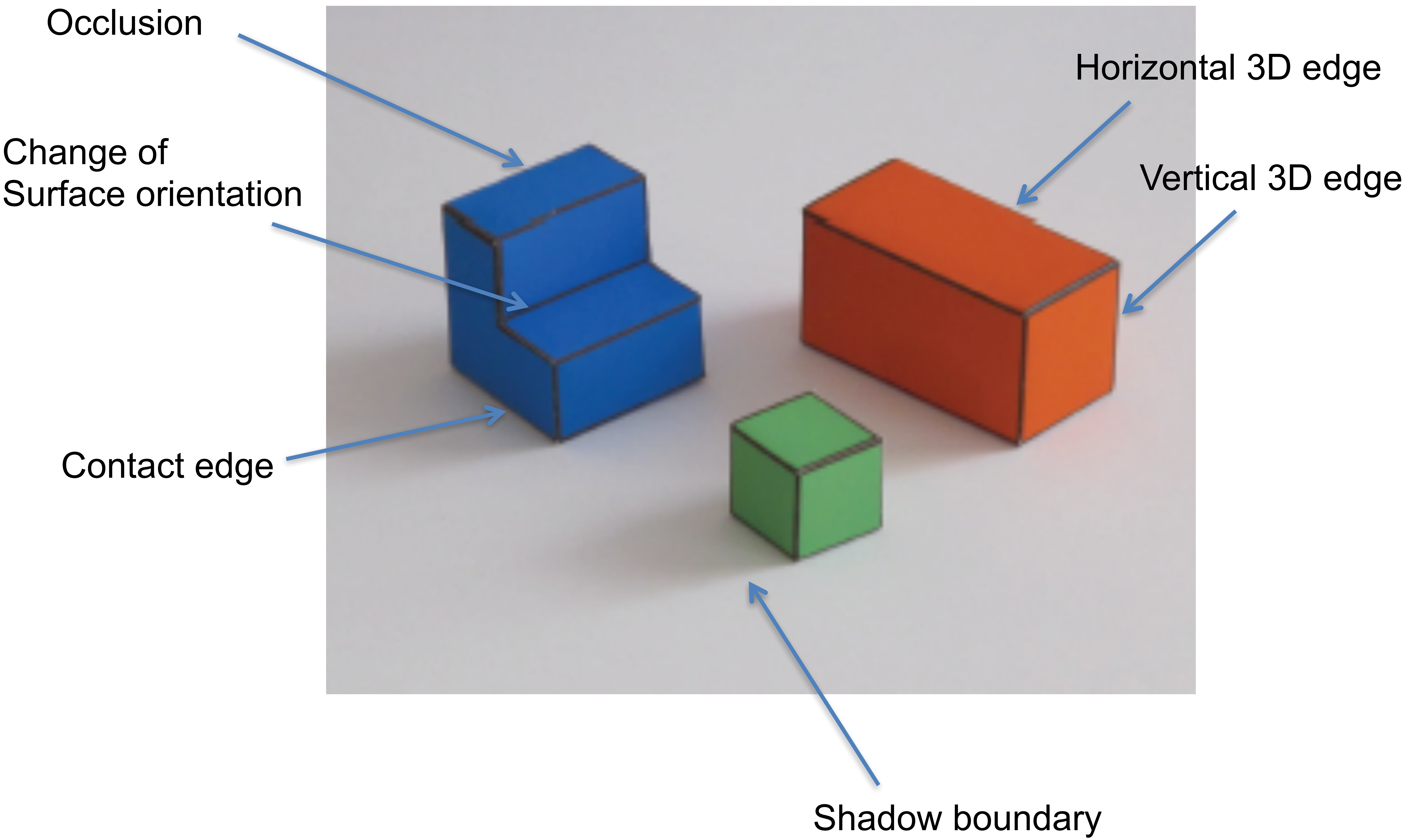
Sinha & Adelson 93

A simple visual system
The input image





Edges


Occlusion
Change of
Surface orientation
Contact edge
Horizontal 3D edge
Vertical 3D edge
Shadow boundary

Finding edges in the image

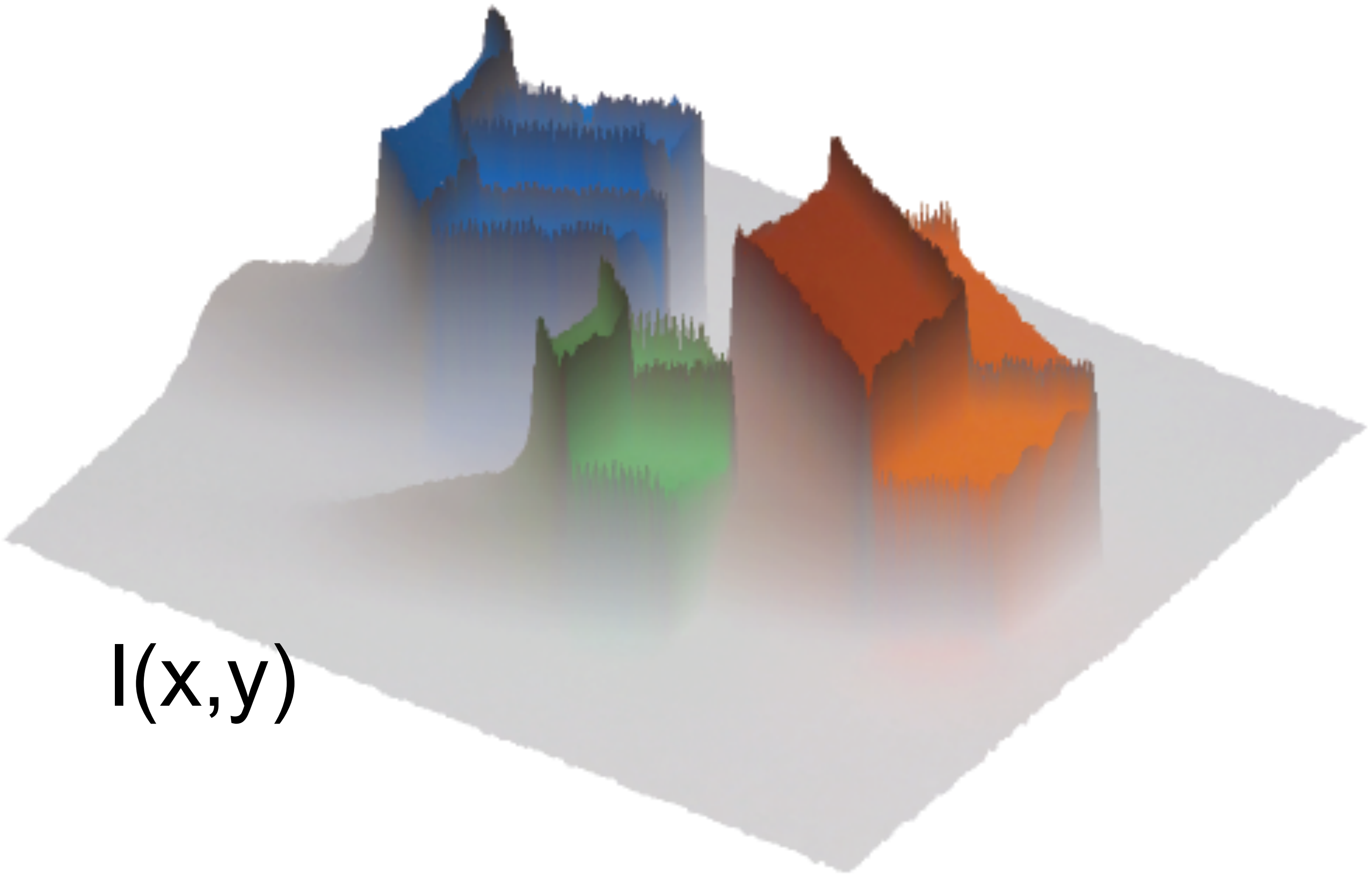


Image gradient:

$$\nabla \mathbf{I} = \left(\frac{\partial \mathbf{I}}{\partial x}, \frac{\partial \mathbf{I}}{\partial y} \right)$$

Approximation image derivative:

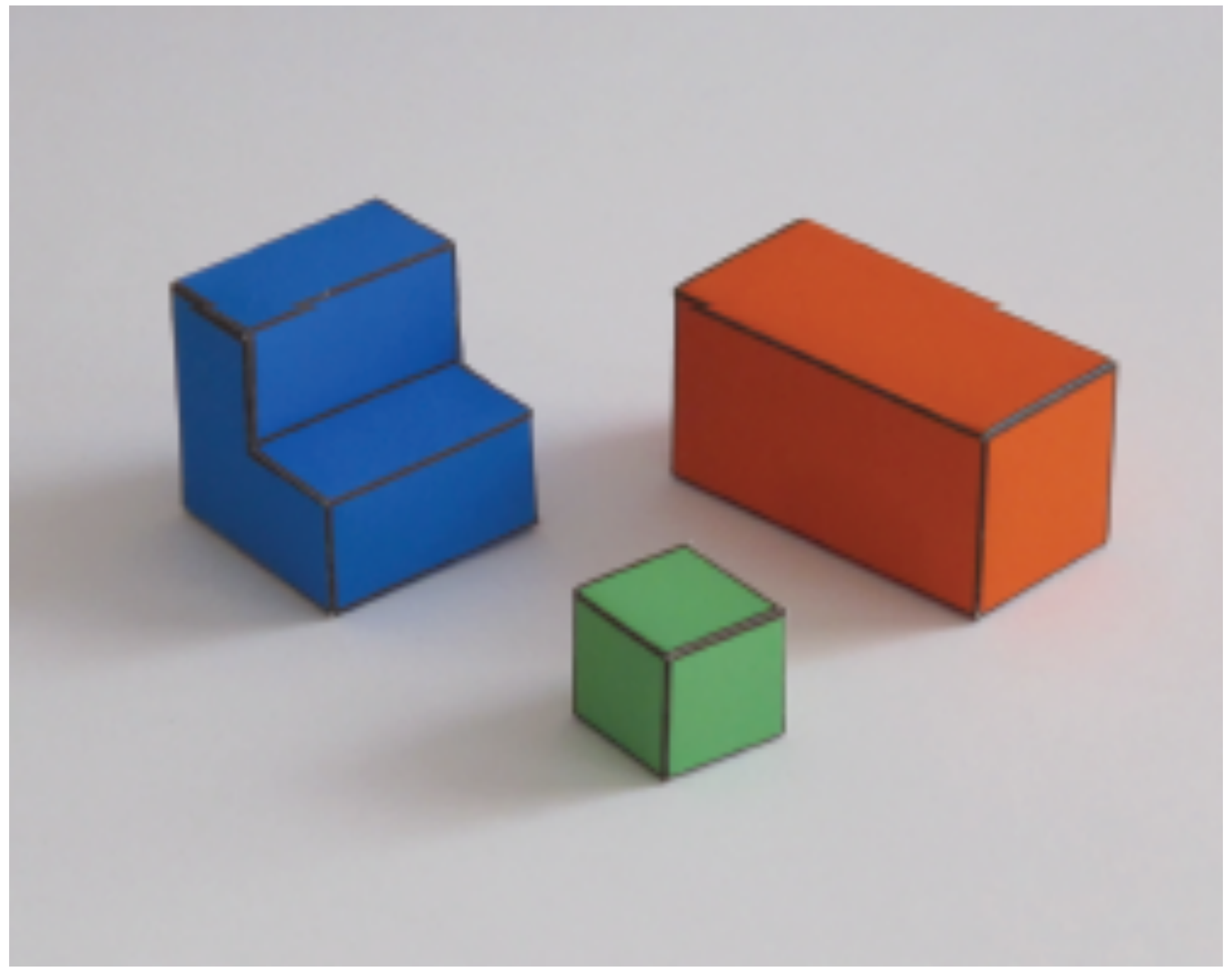
$$\frac{\partial \mathbf{I}}{\partial x} \simeq \mathbf{I}(x,y) - \mathbf{I}(x-1,y)$$

Edge strength $E(x,y) = |\nabla \mathbf{I}(x,y)|$

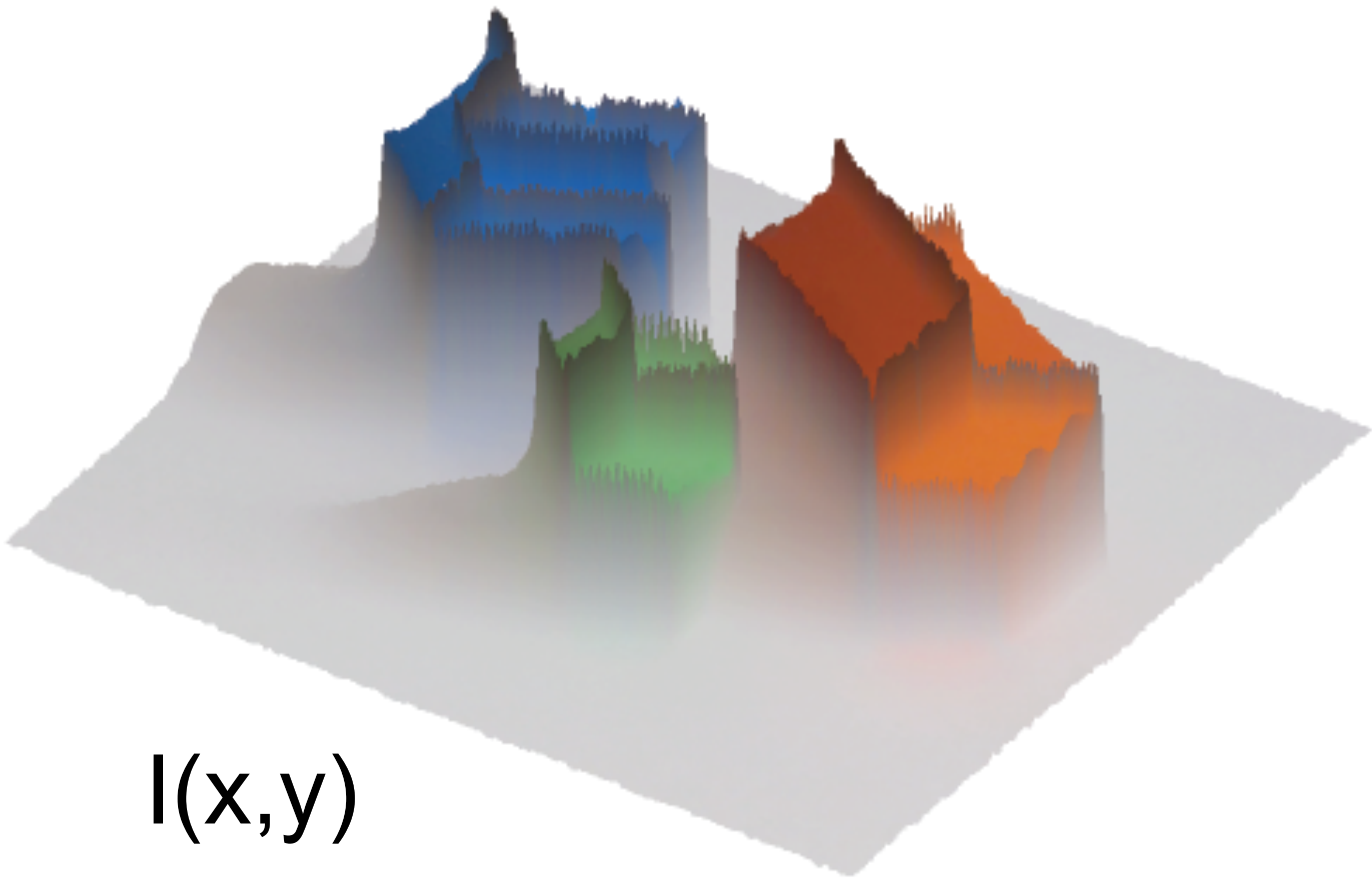
Edge orientation: $\theta(x,y) = \angle \nabla \mathbf{I} = \arctan \frac{\partial \mathbf{I}/\partial y}{\partial \mathbf{I}/\partial x}$

Edge normal: $\mathbf{n} = \frac{\nabla \mathbf{I}}{|\nabla \mathbf{I}|}$

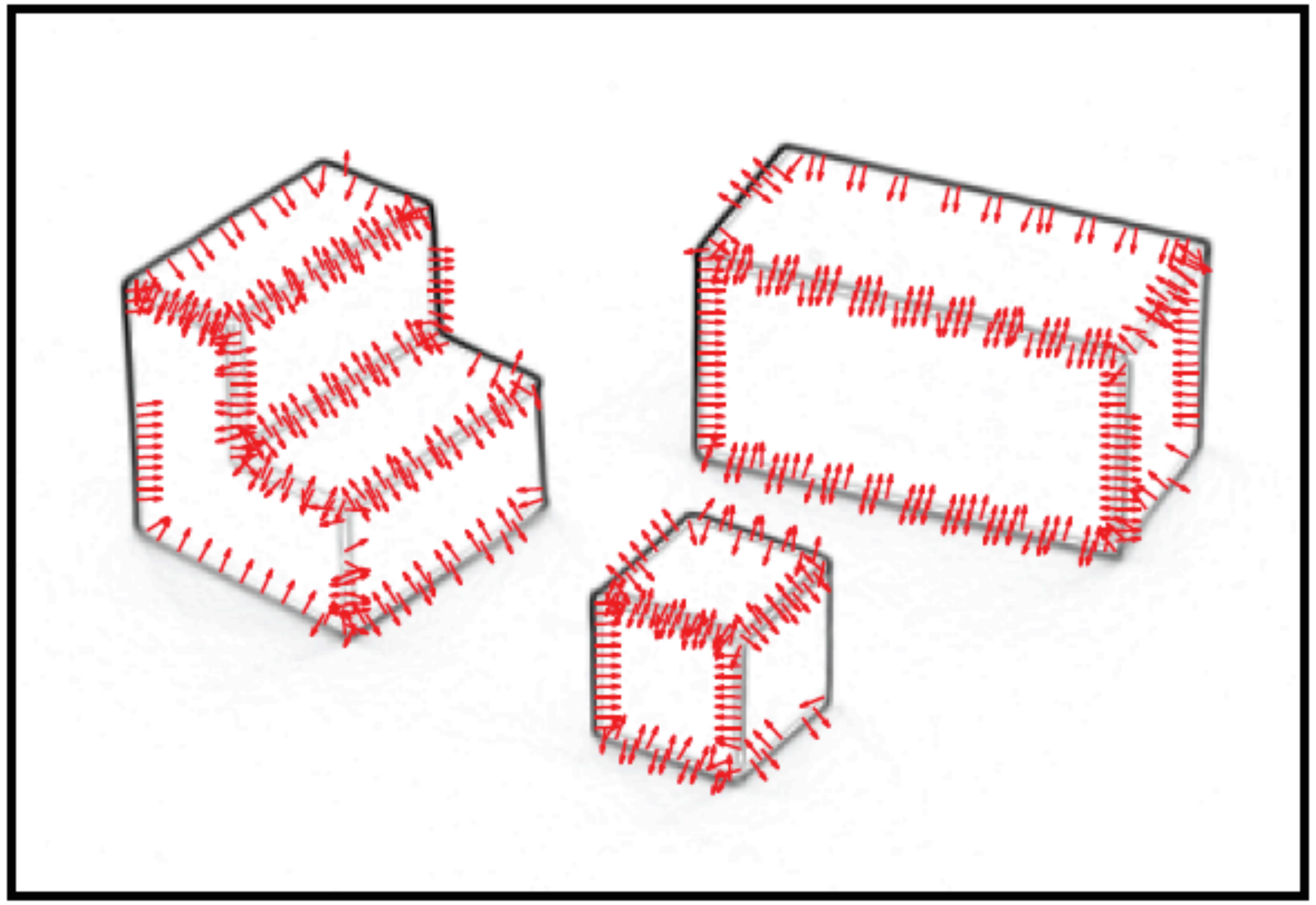
Finding edges in the image



$$\nabla \mathbf{I} = \left(\frac{\partial \mathbf{I}}{\partial x}, \frac{\partial \mathbf{I}}{\partial y} \right) \qquad \mathbf{n} = \frac{\nabla \mathbf{I}}{|\nabla \mathbf{I}|}$$



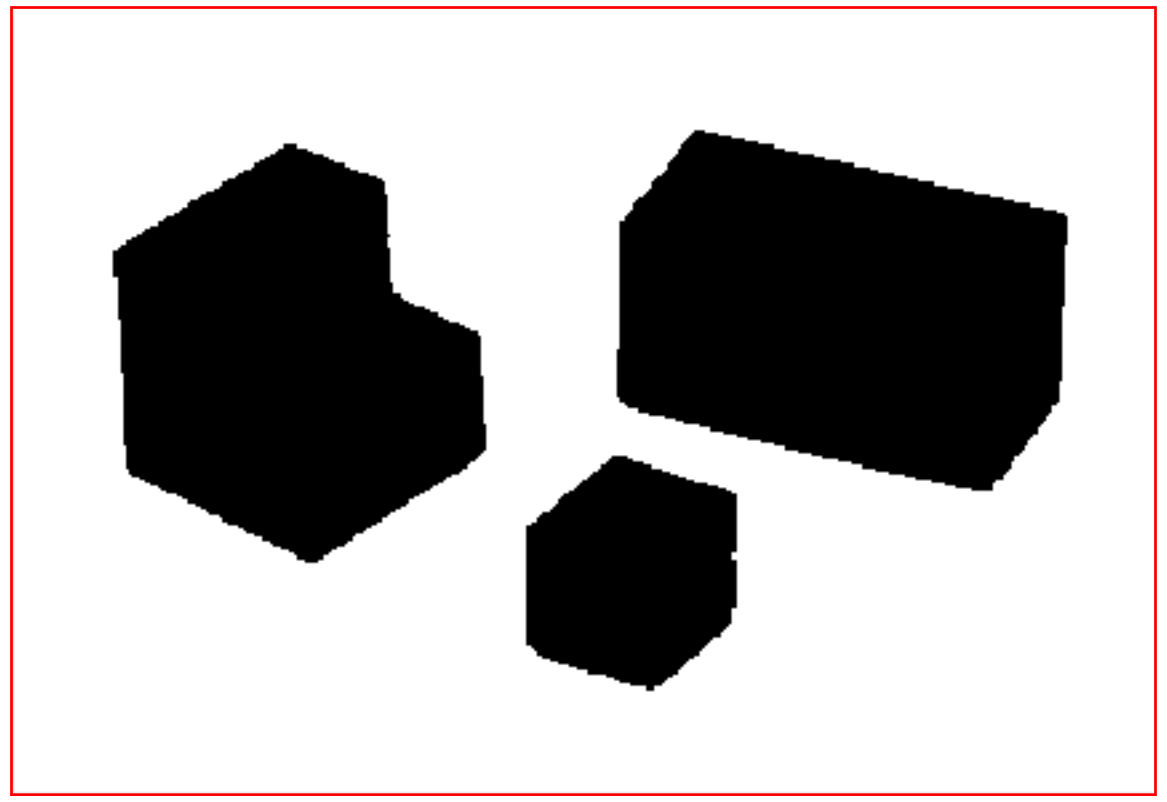
I(x,y)



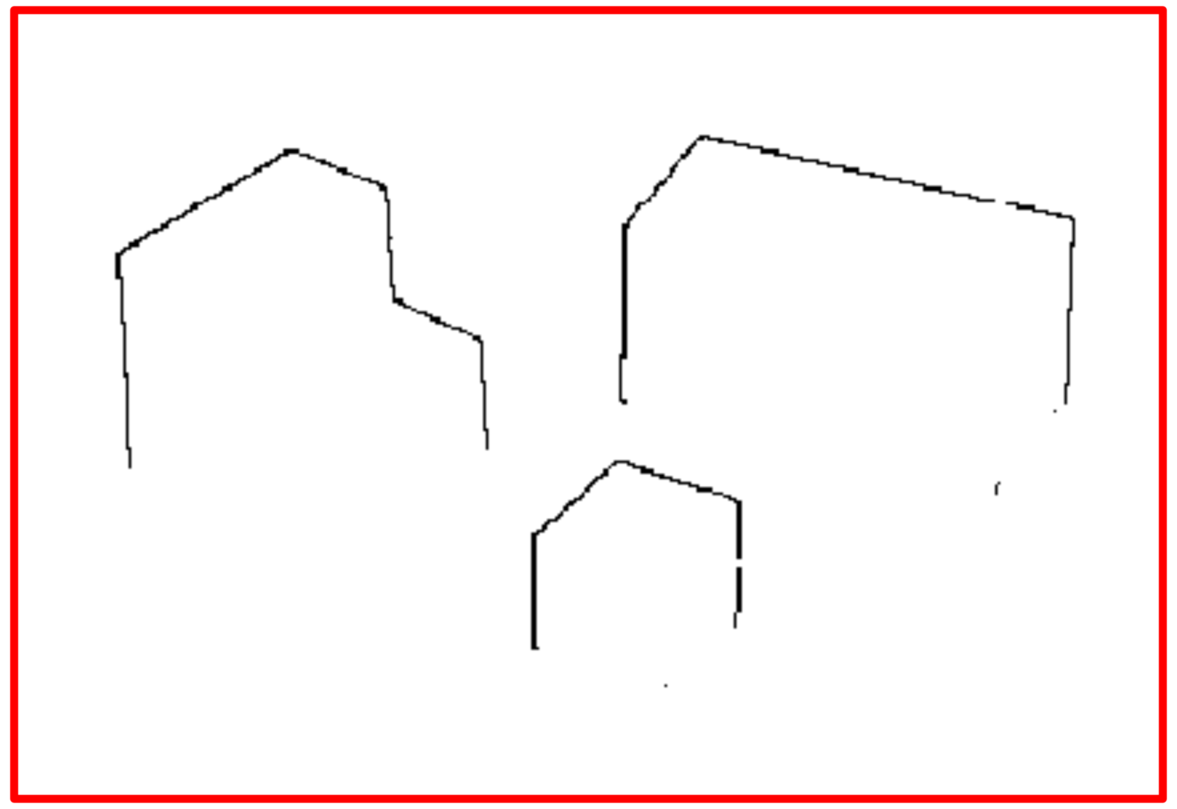
E(x,y) and *n*(x,y)

Edge classification

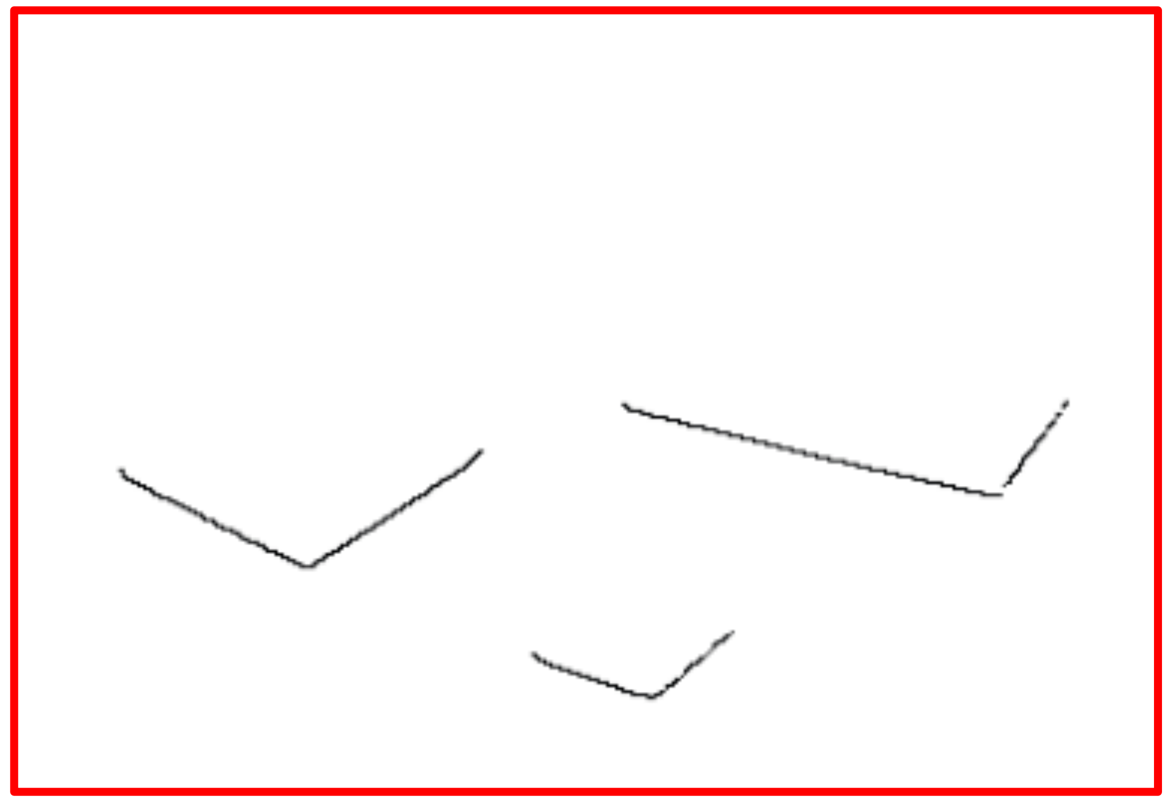
- Figure/ground segmentation
  - Using the fact that objects have color



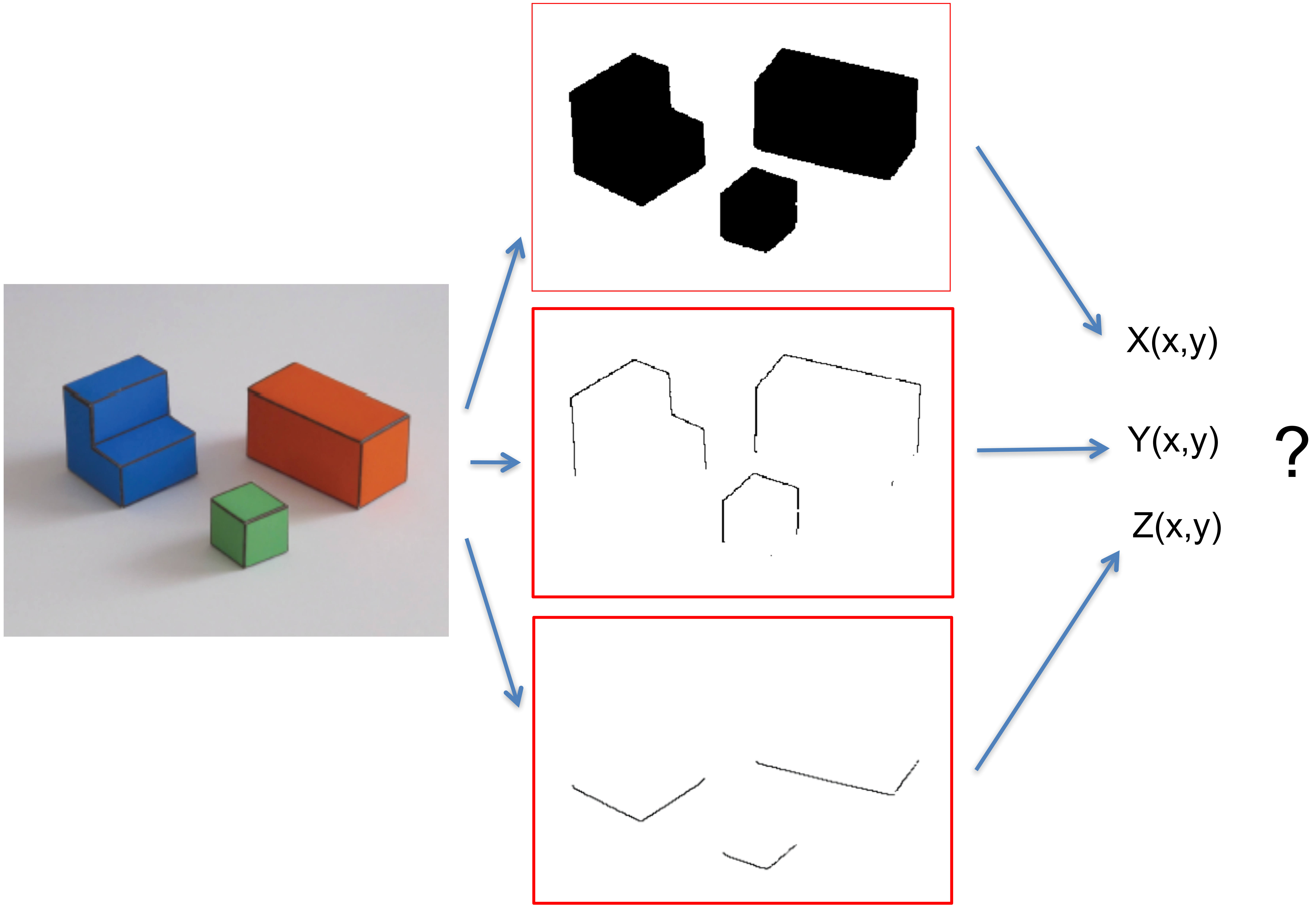
- Occlusion edges
  - Occlusion edges are owned by the foreground



- Contact edges

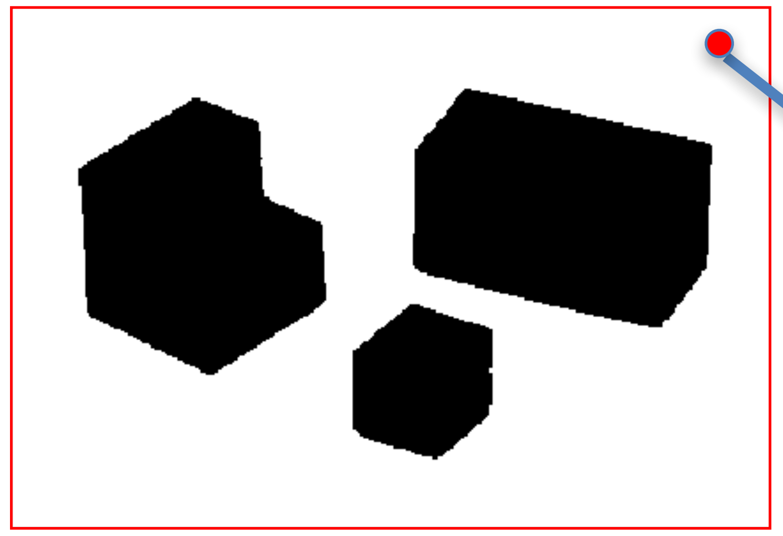


From edges to surface constraints



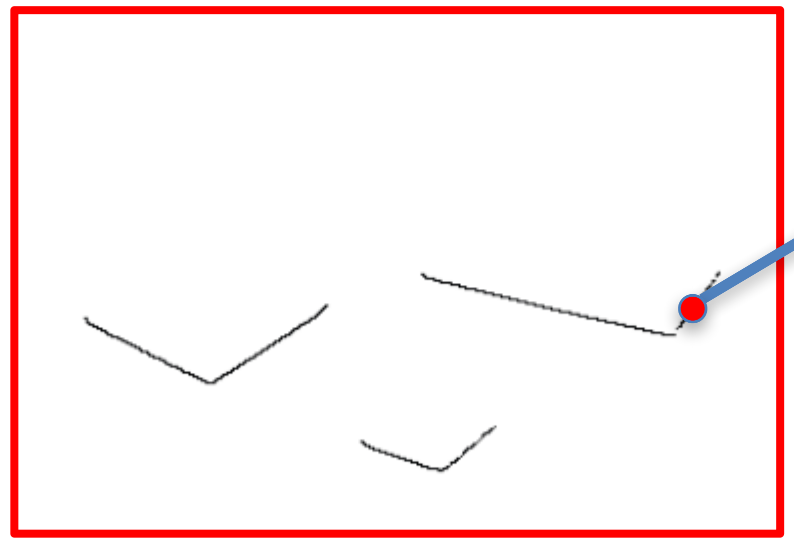
From edges to surface constraints

- Ground



Y(x,y) = 0   if (x,y) belongs to a ground pixel

- Contact edge

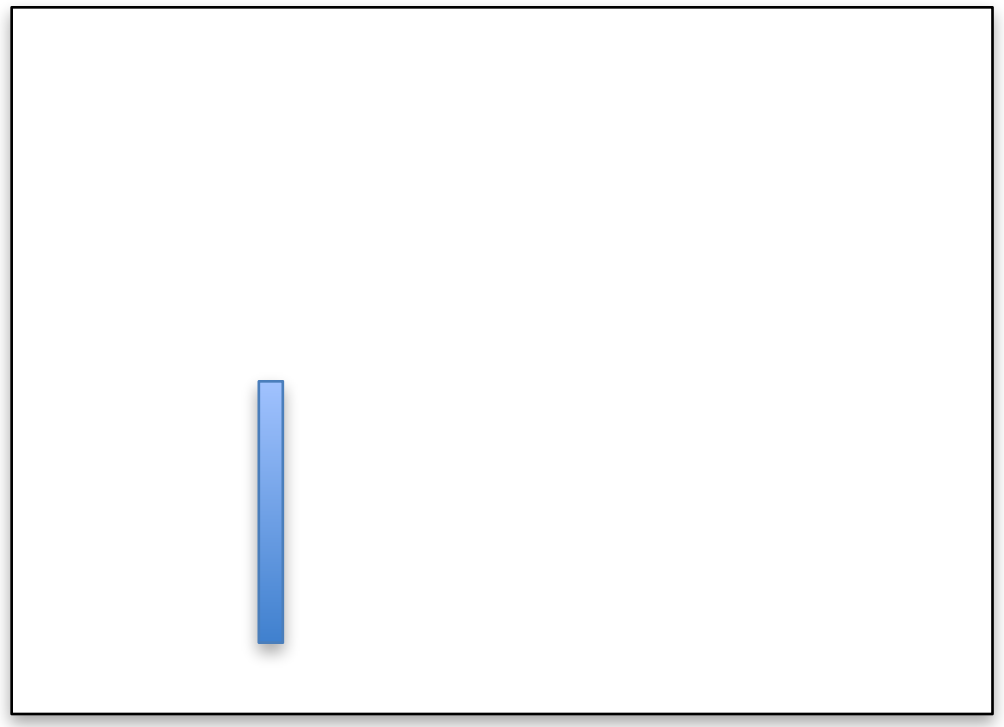


Y(x,y) = 0   if (x,y) belongs to foreground and is a contact edge

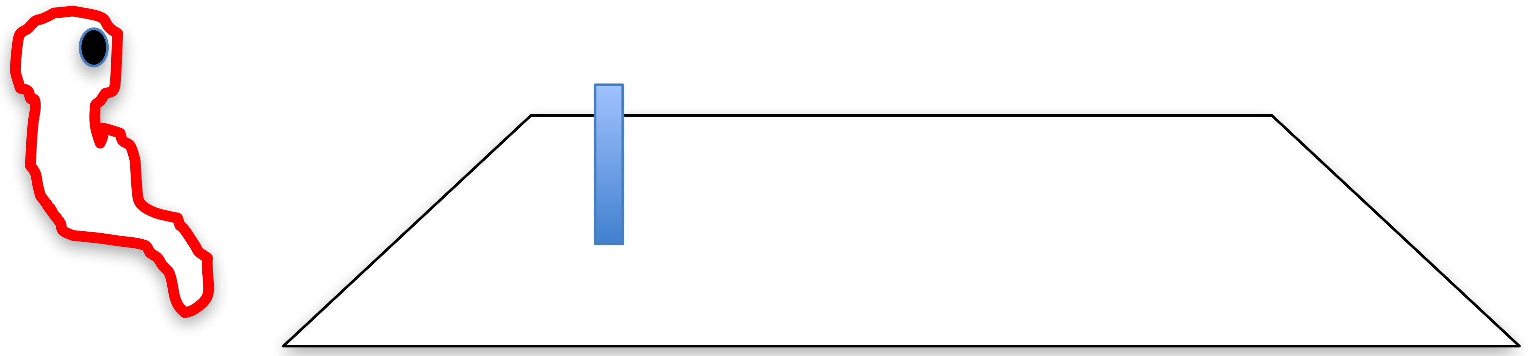
- What happens inside the objects?

… now things get a bit more complicated.

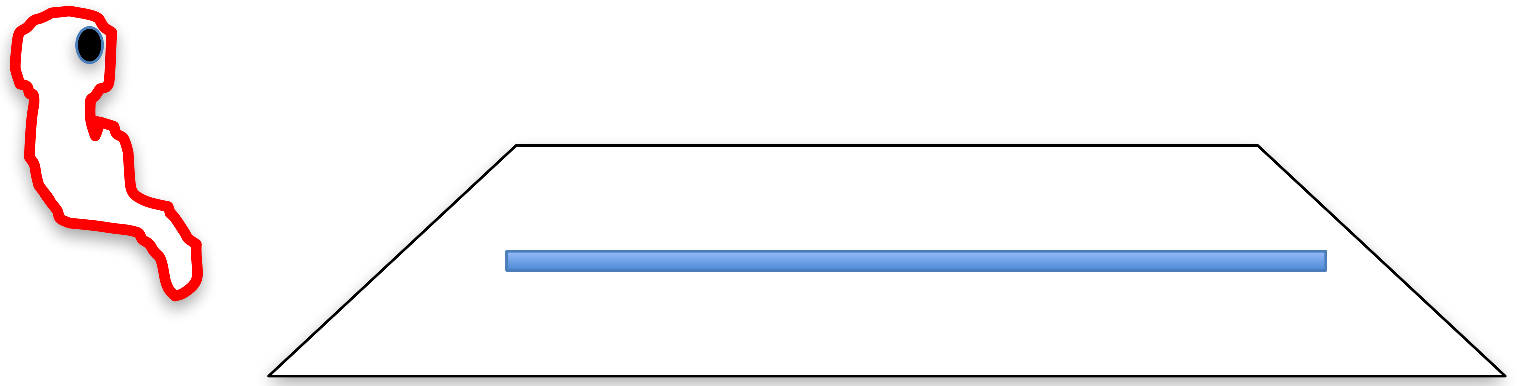
Generic view assumption



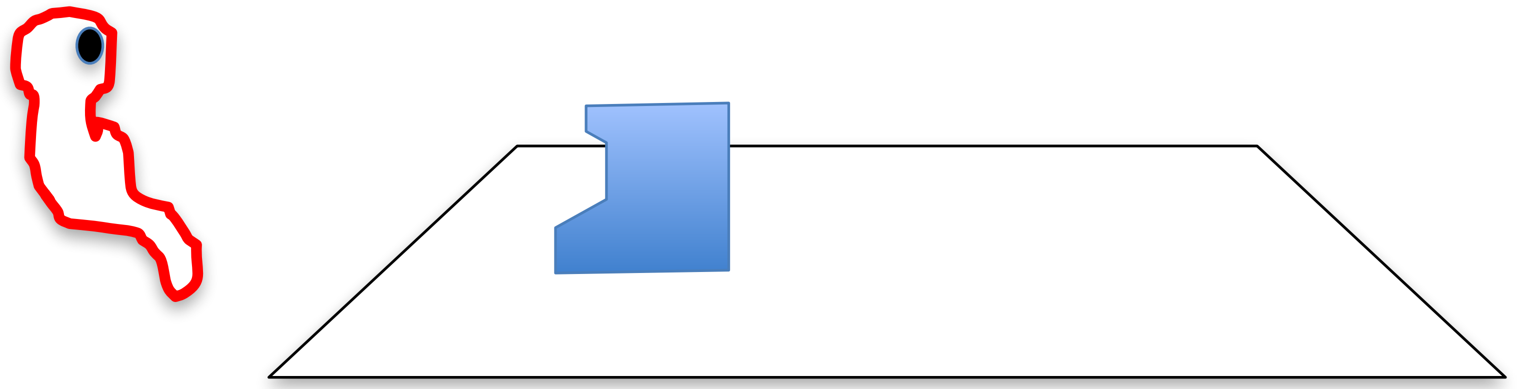
Image



3D world



3D world

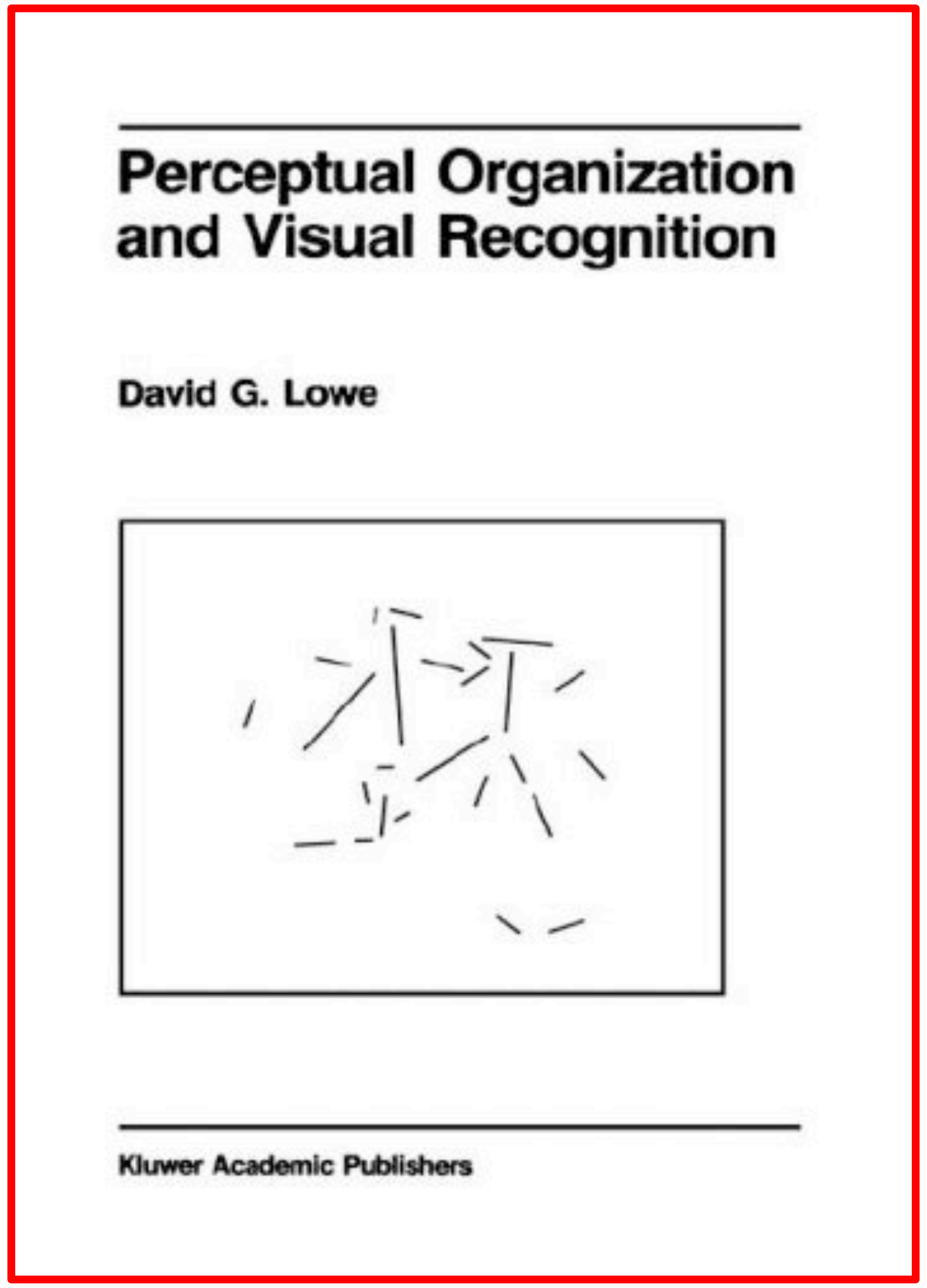


3D world

Generic view assumption: the observer should not assume that he has a special position in the world… The most generic interpretation is to see a vertical line as a vertical line in 3D.

Freeman, 93

Non-accidental properties



D. Lowe, 1985

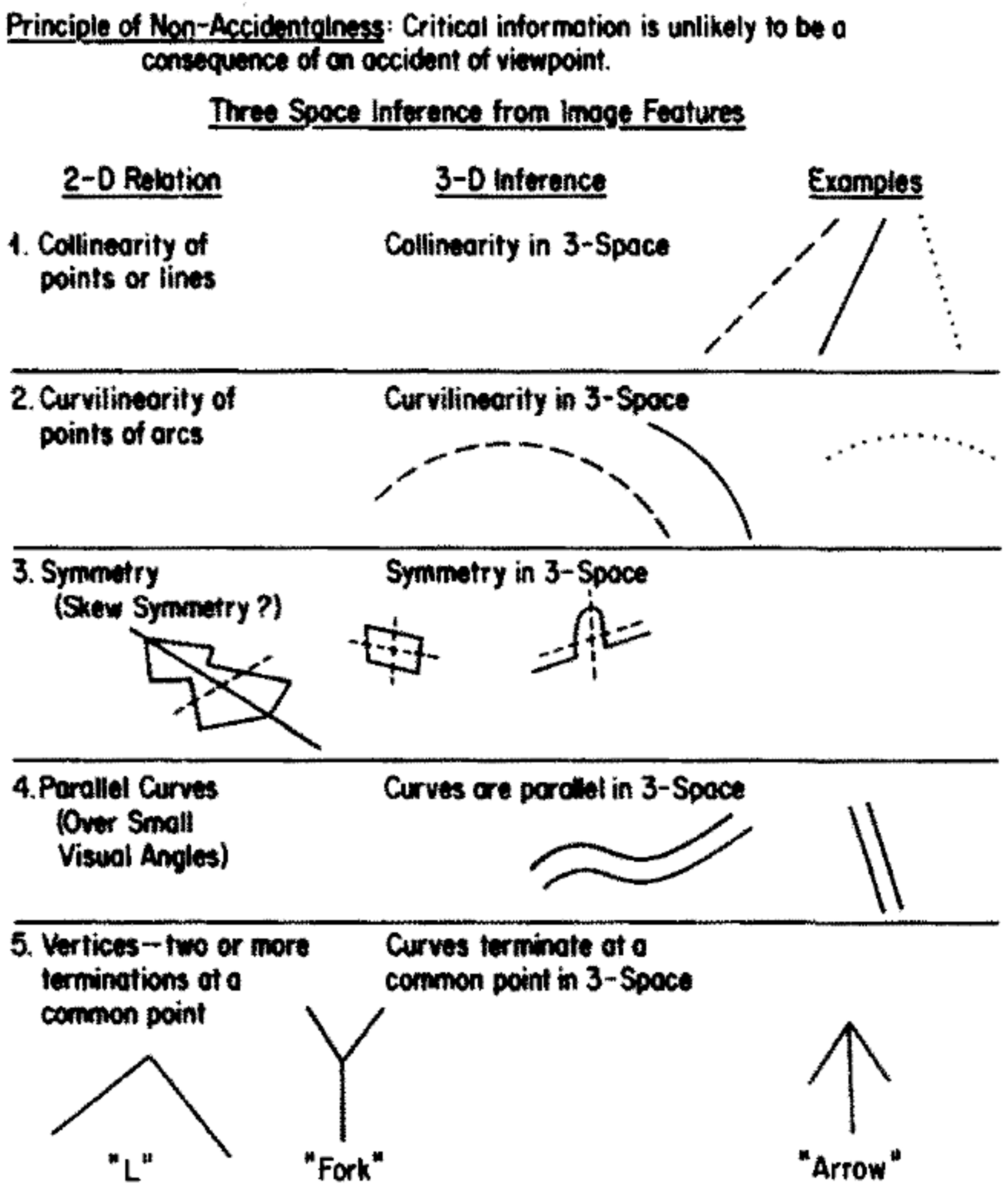
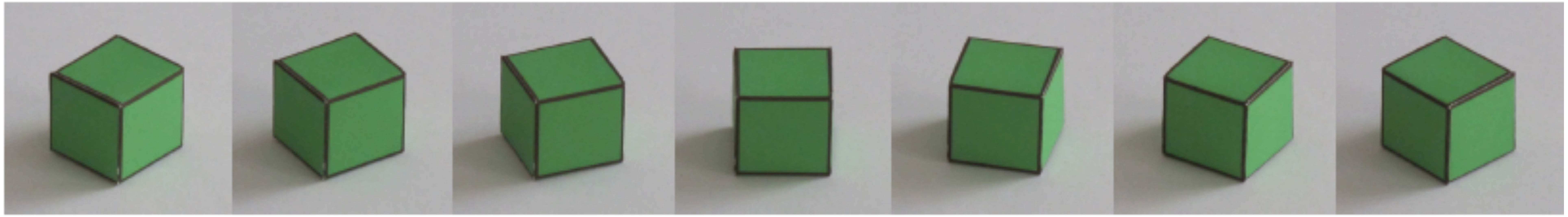


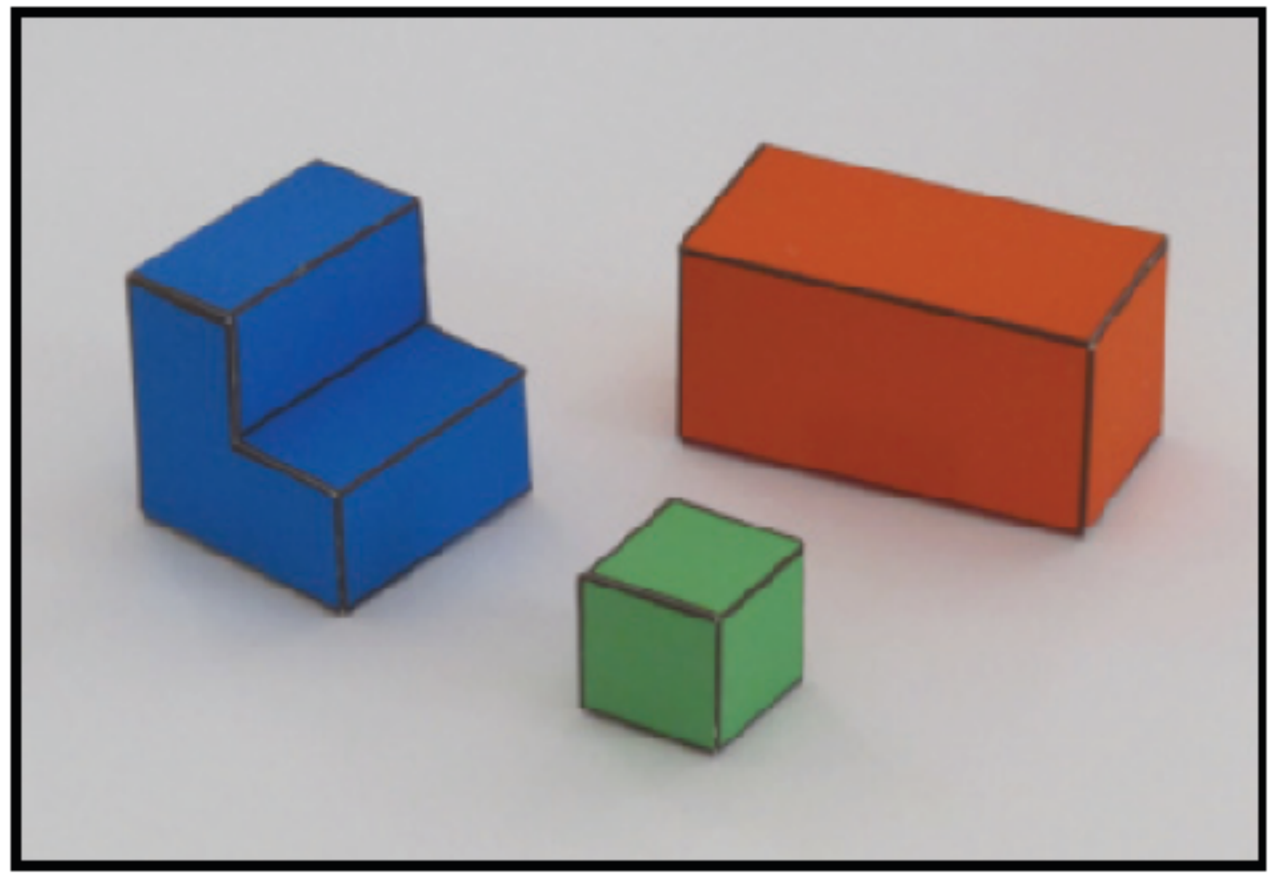
Figure 4. Five nonaccidental relations. (From Figure 5.2, *Perceptual organization and visual recognition* [p. 77] by David Lowe. Unpublished doctorial dissertation, Stanford University. Adapted by permission.)

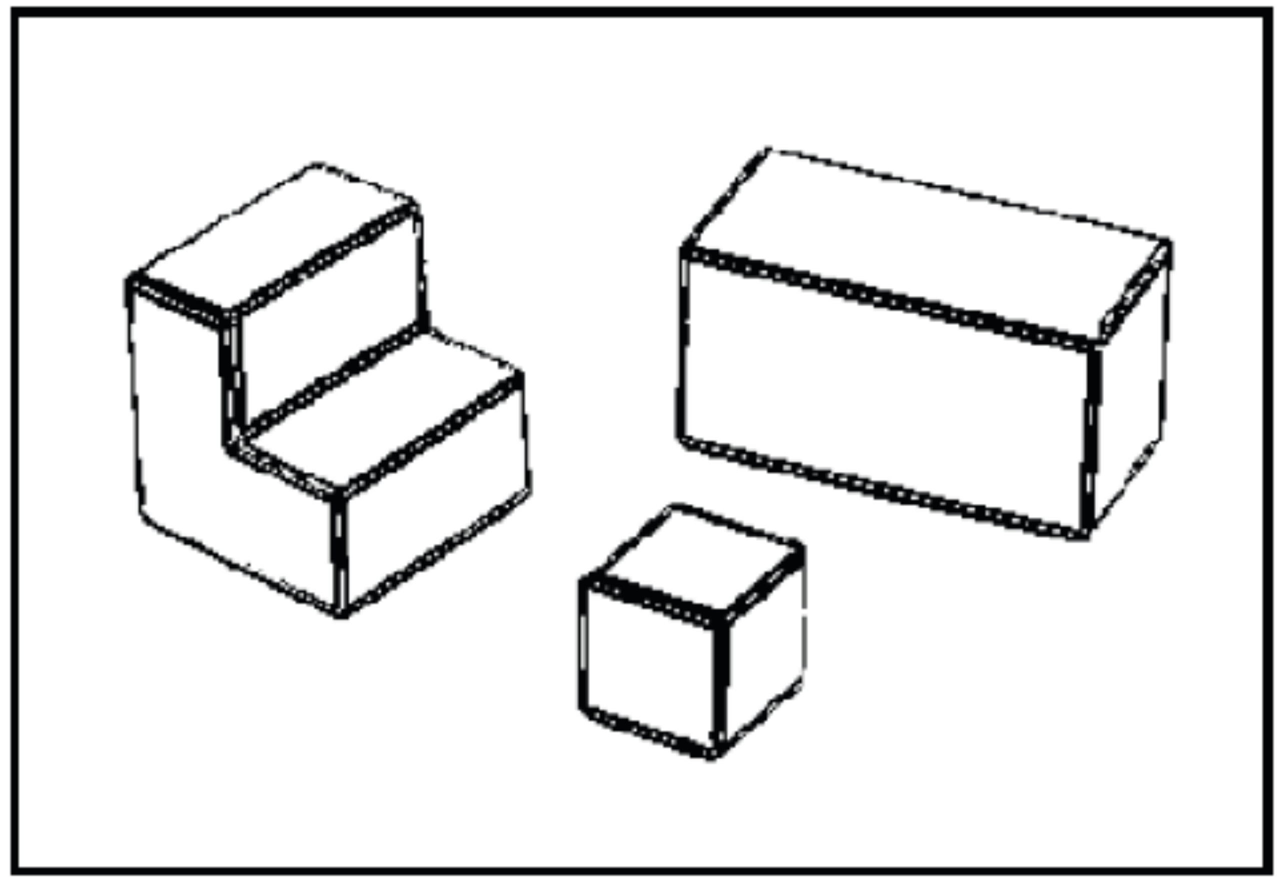
Biederman_RBC_1987

Non-accidental properties in the simple world

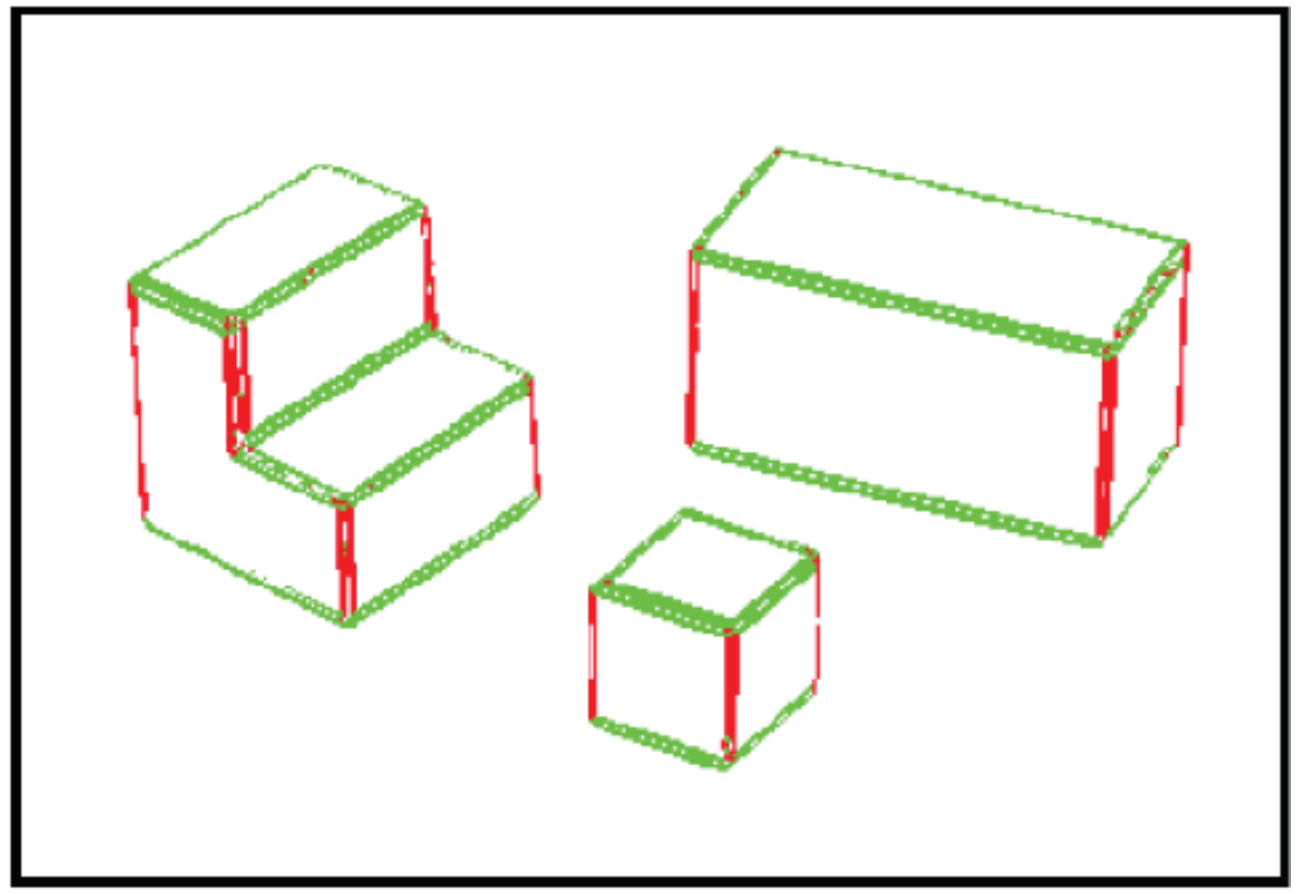


generic generic generic accidental generic generic generic





Using E(x,y)

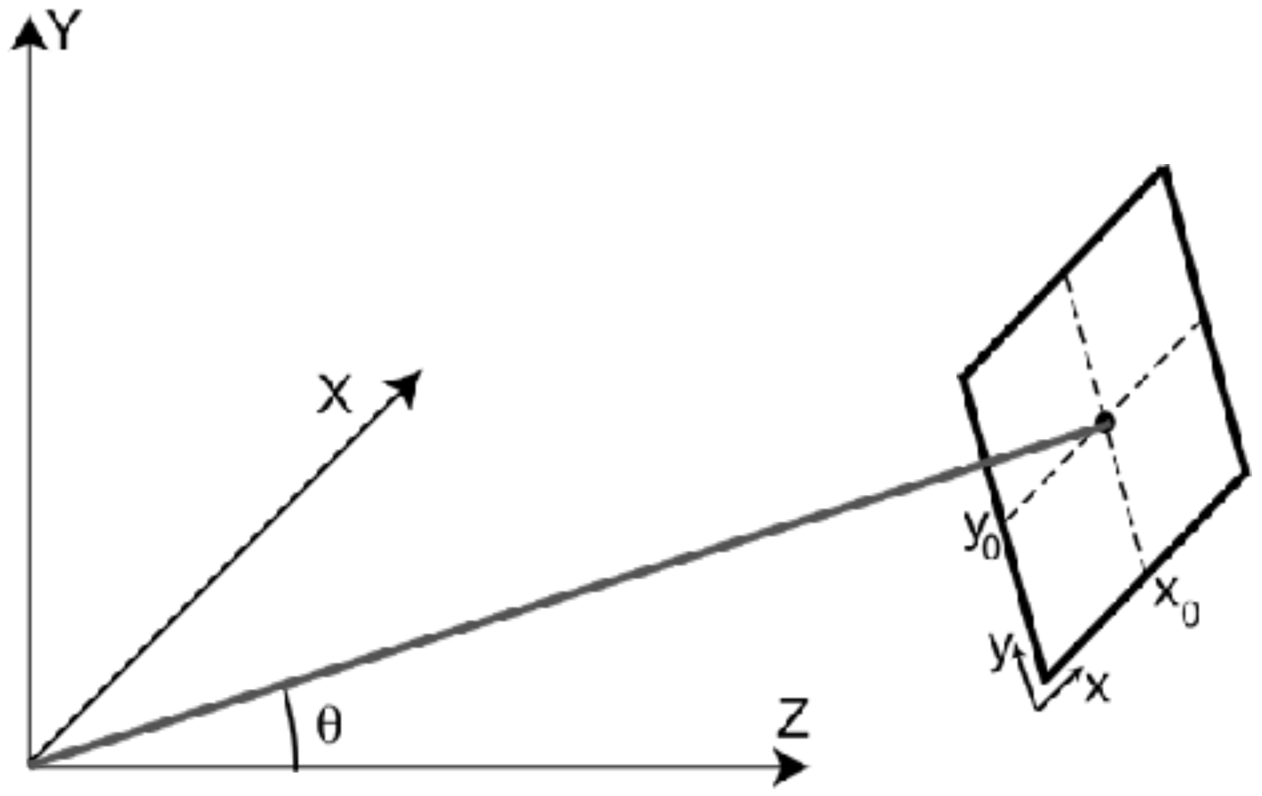


Using θ(x,y)

From edges to surface constraints

How can we relate the information in the pixels with 3D surfaces in the world?

- Vertical edges



World coordinates

$x = X + x_0$

$y = \cos(\theta)\, Y - \sin(\theta)\, Z + y_0$

image coordinates

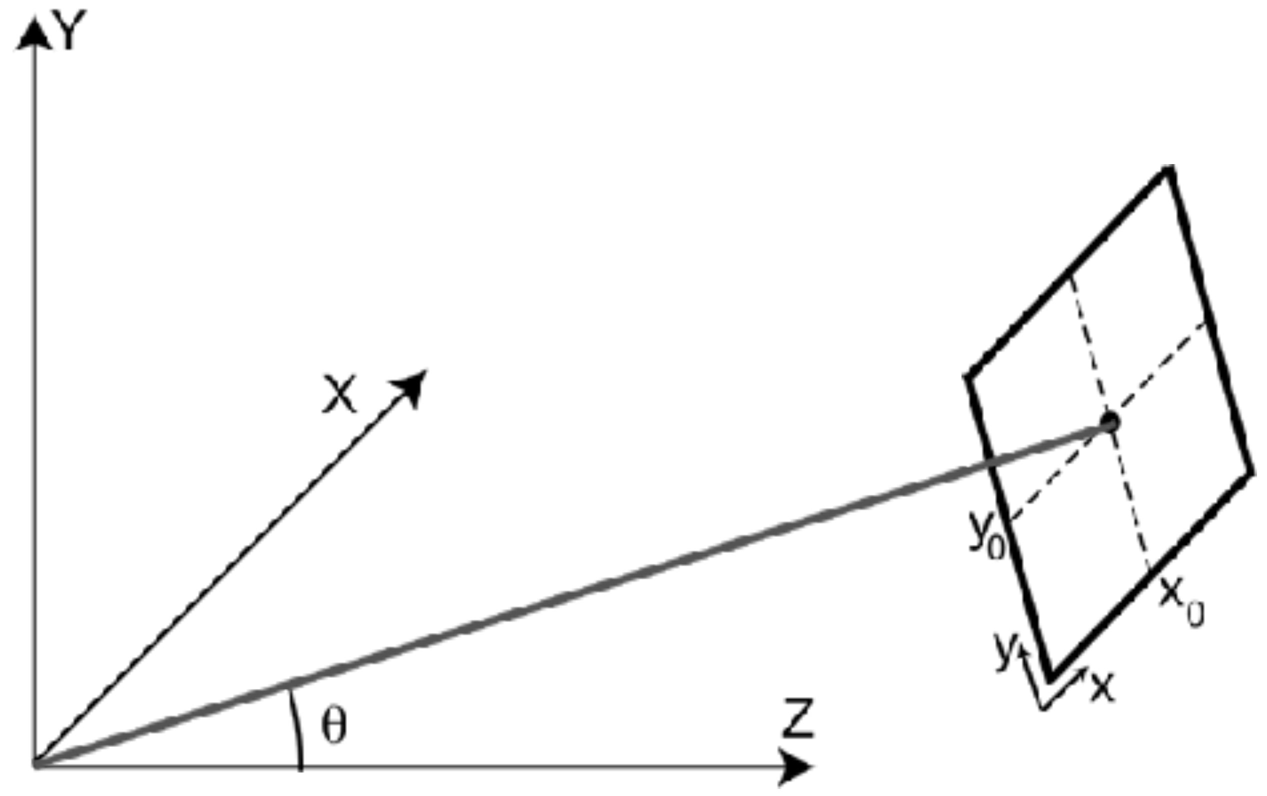
Given the image, what can we say about X, Y and Z in the pixels that belong to a vertical edge?

Z = constant along the edge

$$\partial Y / \partial y = 1 / \cos(\theta)$$

From edges to surface constraints

- Horizontal edges



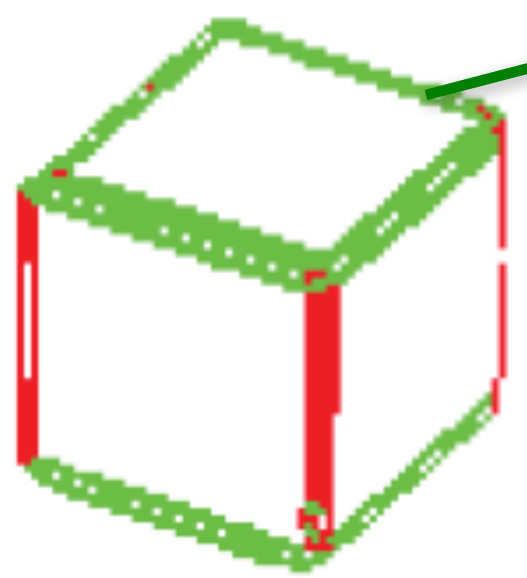
World coordinates

$x = X + x_0$

$y = \cos(\theta)\, Y - \sin(\theta)\, Z + y_0$

image coordinates

Given the image, what can we say about X, Y and Z in the pixels that belong to an horizontal 3D edge?



Y = constant along the edge

$$\partial Y / \partial \mathbf{t} = 0$$

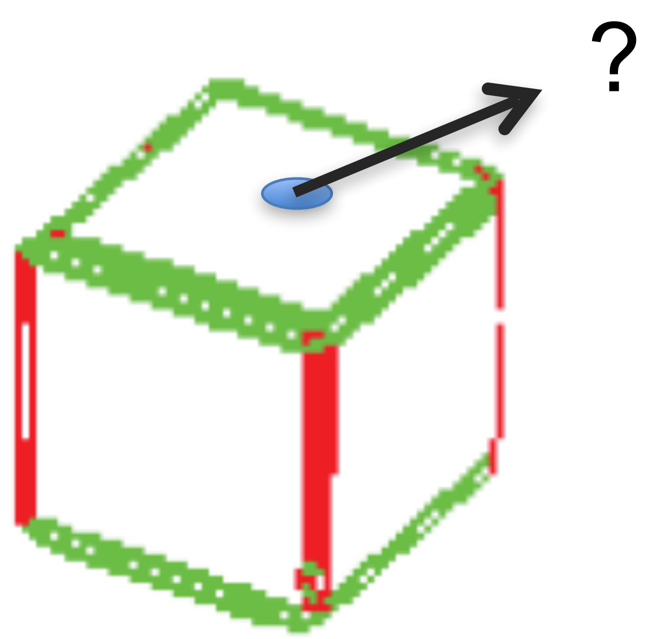
Where **t** is the vector parallel to the edge

$$\mathbf{t} = (-n_y, n_x)$$

$$\partial Y / \partial \mathbf{t} = -n_y \partial Y / \partial x + n_x \partial Y / \partial y$$

From edges to surface constraints

- What happens where there are no edges?



Assumption of planar faces:

$$\partial^2 Y/\partial x^2 = 0$$
$$\partial^2 Y/\partial y^2 = 0$$
$$\partial^2 Y/\partial y \partial x = 0$$

Information has to be propagated from the edges

A simple inference scheme

All the constraints are linear

$Y(x,y) = 0$ if (x,y) belongs to a ground pixel

$$\partial Y/\partial y = 1/\cos(\theta)$$

if (x,y) belongs to a vertical edge

$$\partial Y/\partial t = 0$$

if (x,y) belongs to an horizontal edge

$$\begin{aligned} \partial^2 Y/\partial x^2 &= 0 \\ \partial^2 Y/\partial y^2 &= 0 \\ \partial^2 Y/\partial y \partial x &= 0 \end{aligned}$$

if (x,y) is not on an edge

A similar set of constraints could be derived for Z

Discrete approximation

We can transform every differential constrain into a discrete linear constraint on Y(x,y)

Y(x,y)

111	115	113	111	112	111	112	111
135	138	137	139	145	146	149	147
163	168	188	196	206	202	206	207
180	184	206	219	202	200	195	193
189	193	214	216	104	79	83	77
191	201	217	220	103	59	60	68
195	205	216	222	113	68	69	83
199	203	223	228	108	68	71	77

$$\frac{dY}{dx} \approx Y(x,y) - Y(x-1,y)$$

-1	1

A slightly better approximation

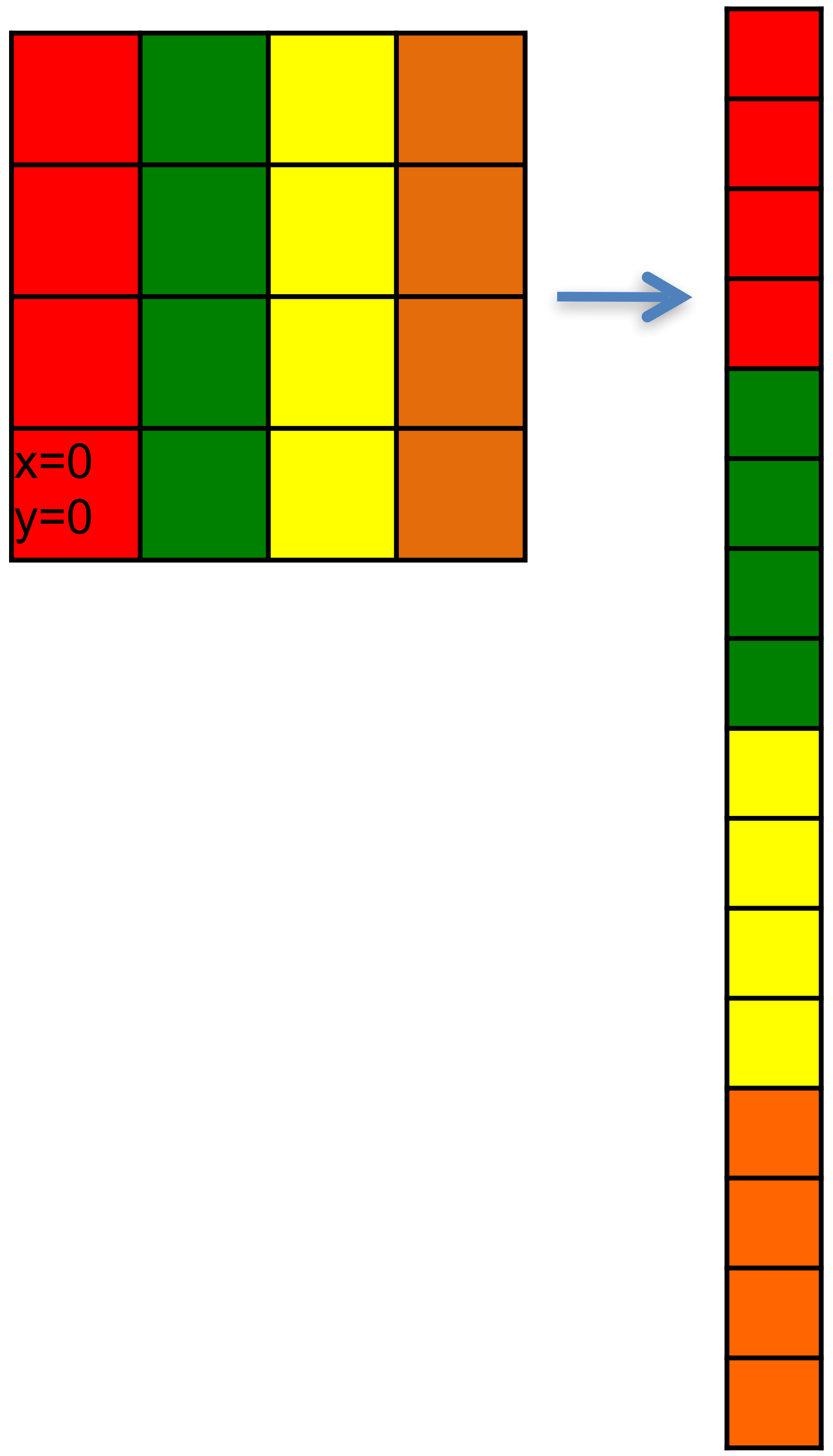
(it is symmetric, and it averages horizontal derivatives over 3 vertical locations)

-1	0	1
-2	0	2
-1	0	1

Discrete approximation

Transform the “image” Y(x,y) into a column vector:

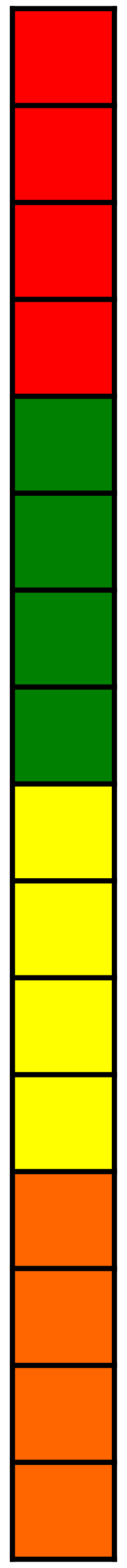
Y(x,y)



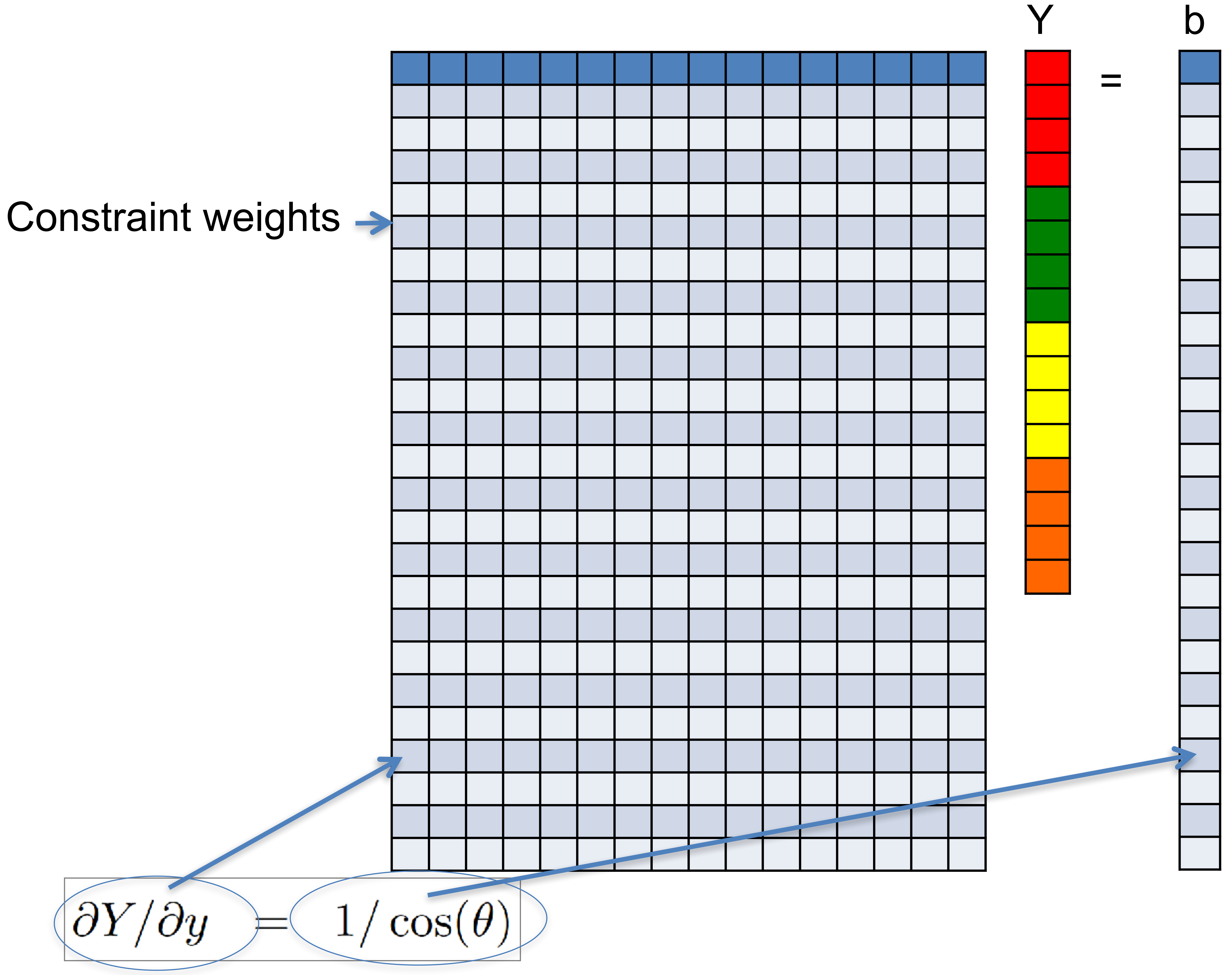
x=2, y=2

$$\frac{dY}{dx} \approx Y(x,y) - Y(x-1,y) = Y(2,2) - Y(1,2) =$$

0	0	0	0	0	-1	0	0	0	1	0	0	0	0	0	0



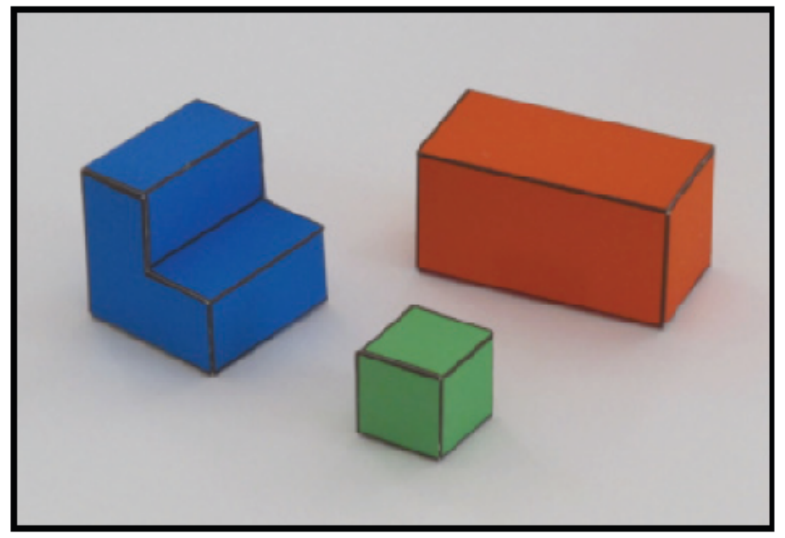
A simple inference scheme



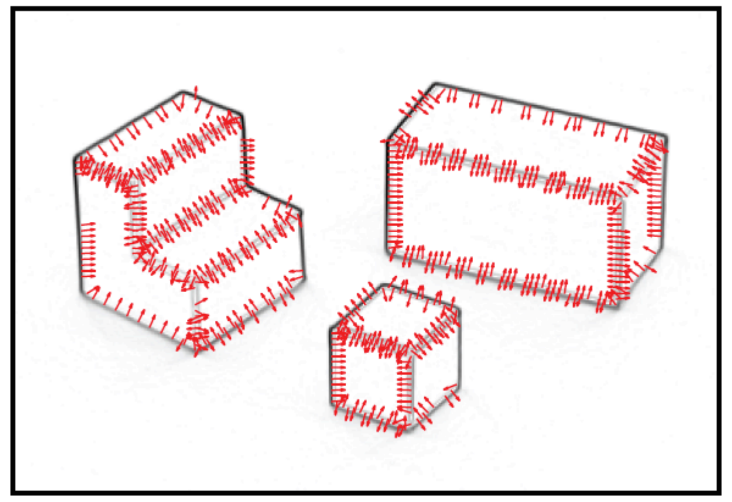
A Y = b

$Y = (A^T A)^{-1} A^T b$

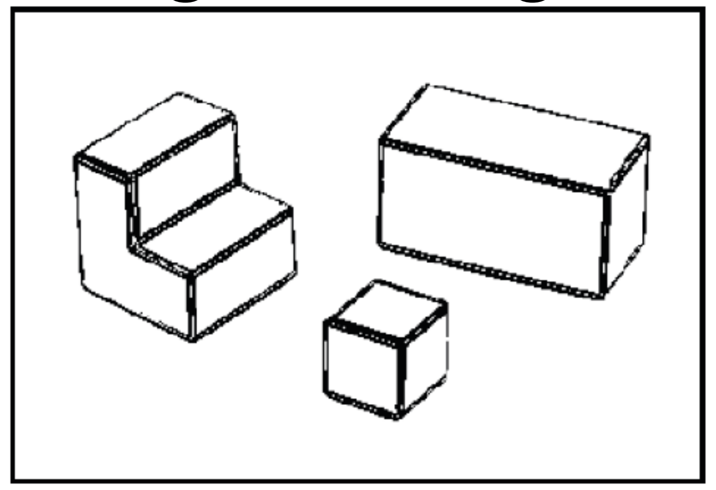
Results



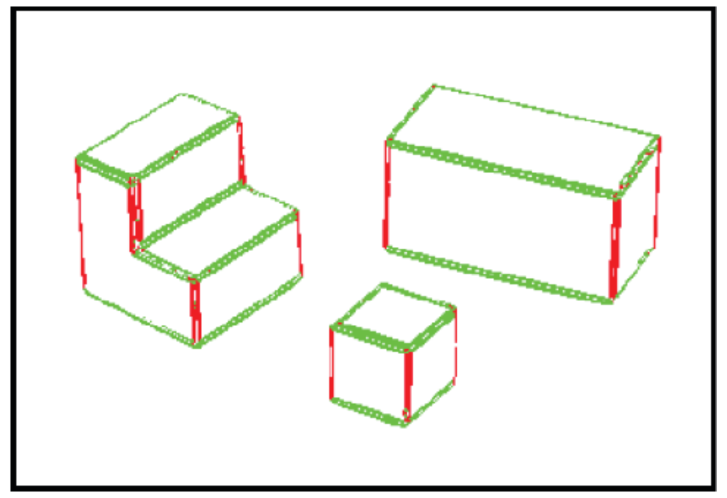
Edge normals



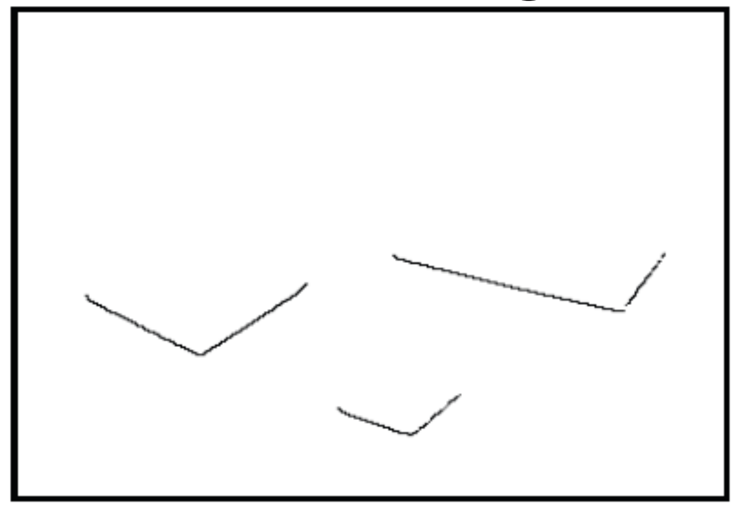
Edge strength



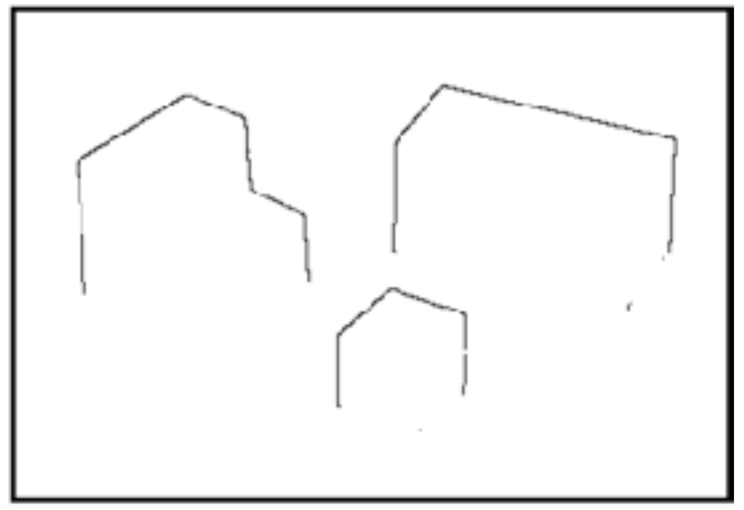
3D orientation

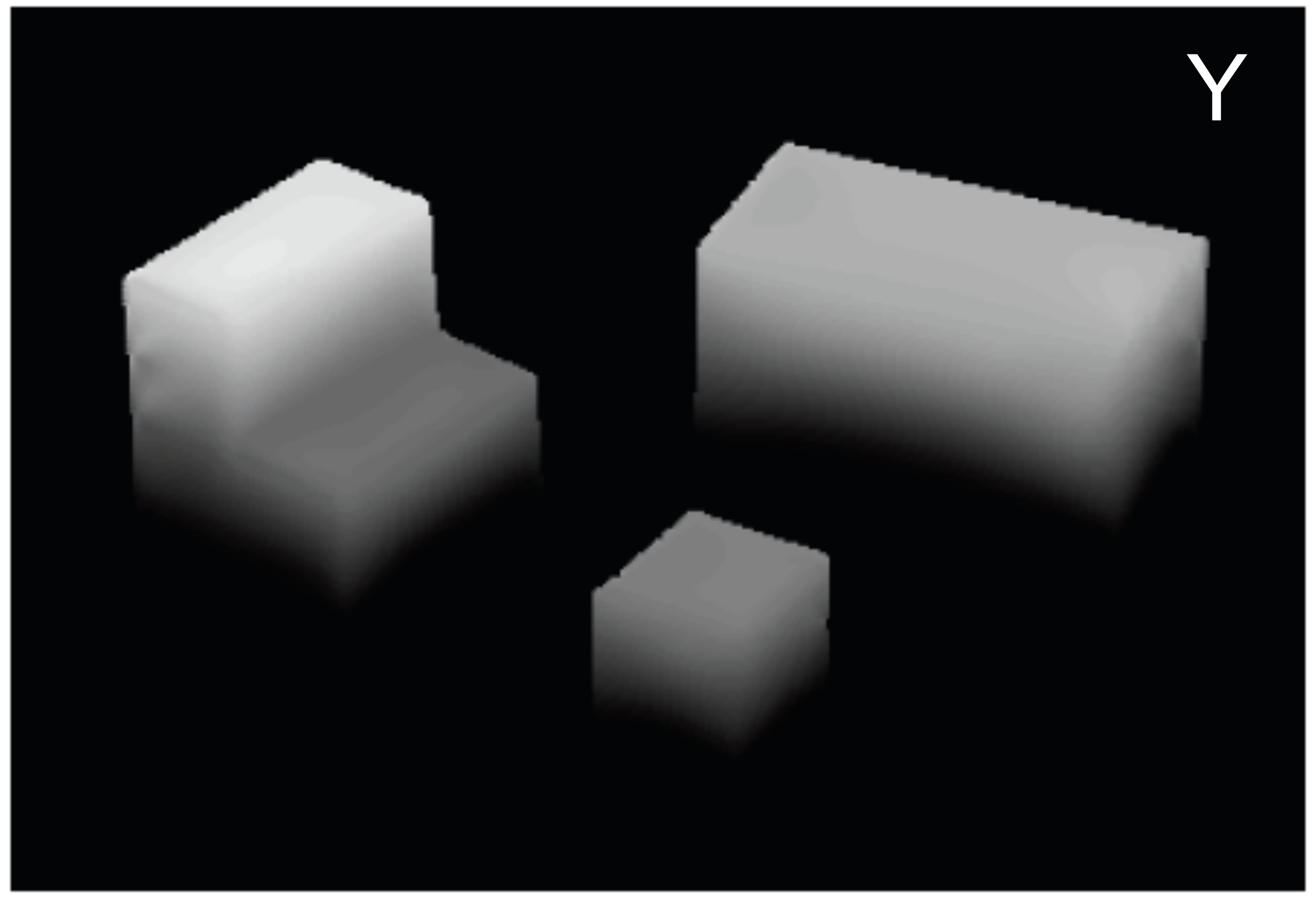


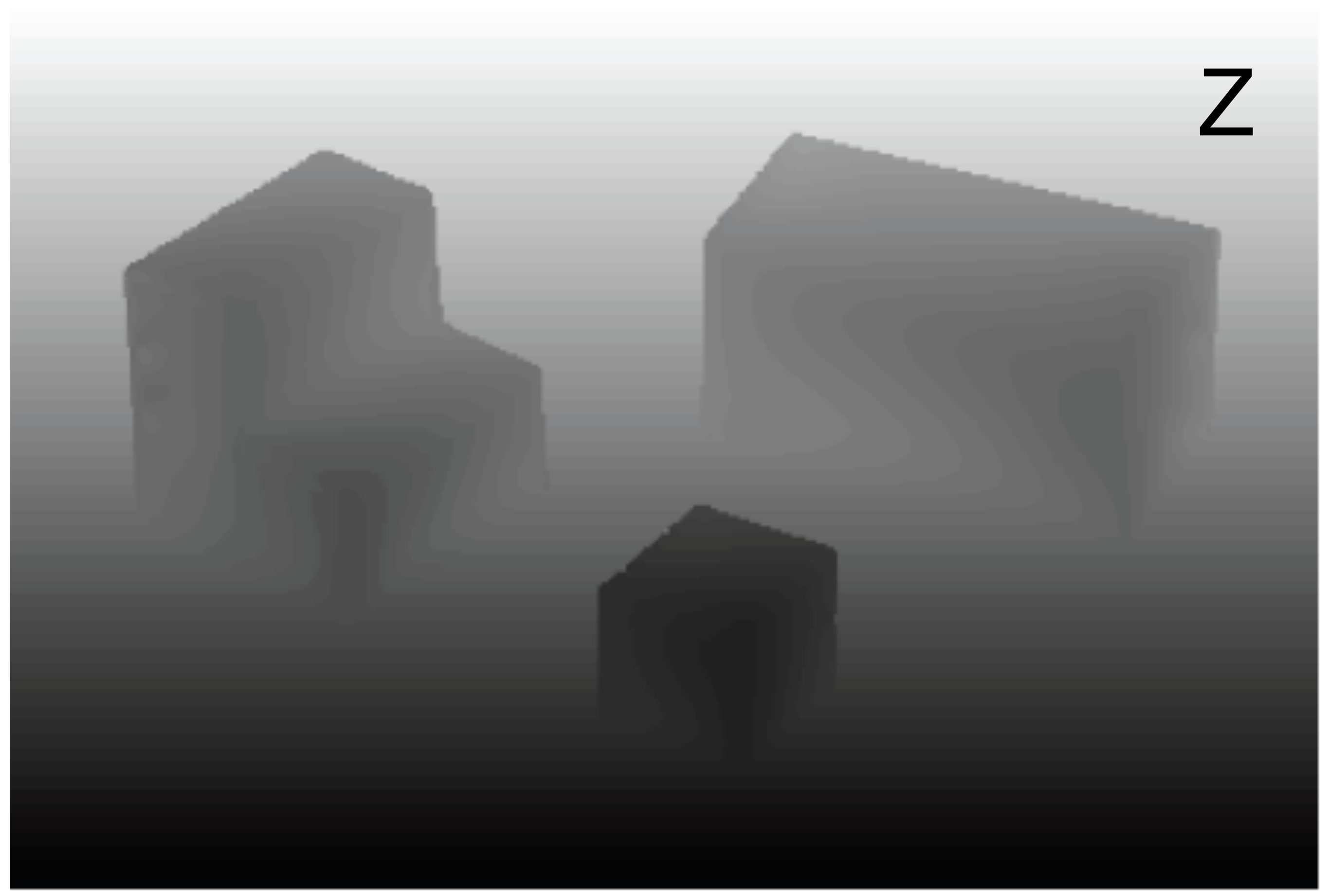
Contact edges



Depth discontinuities

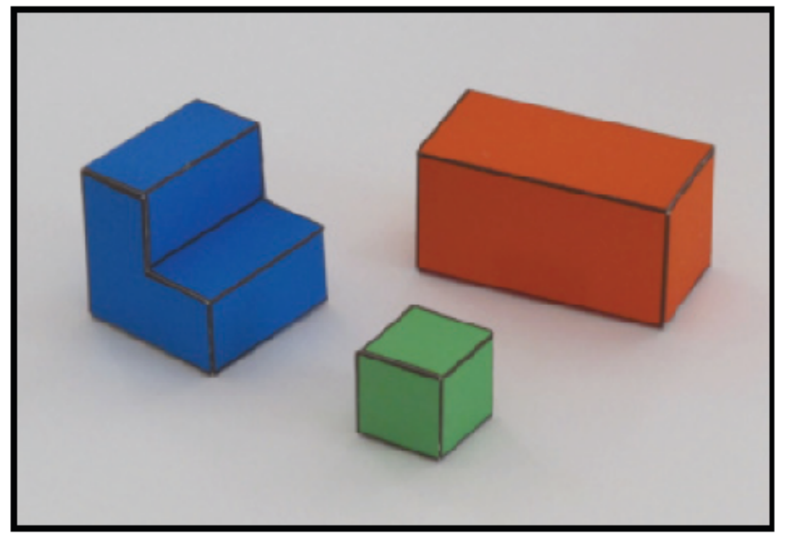




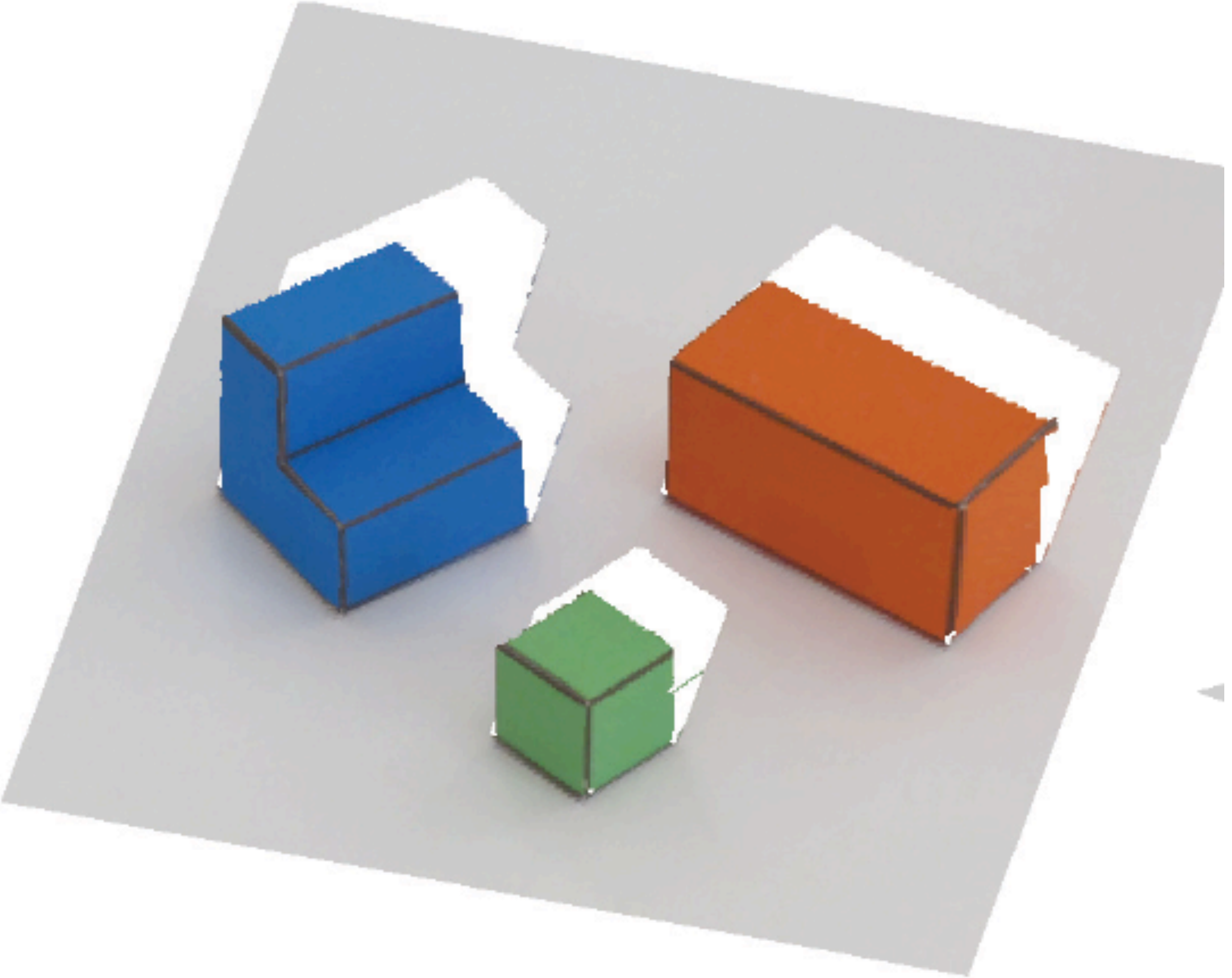


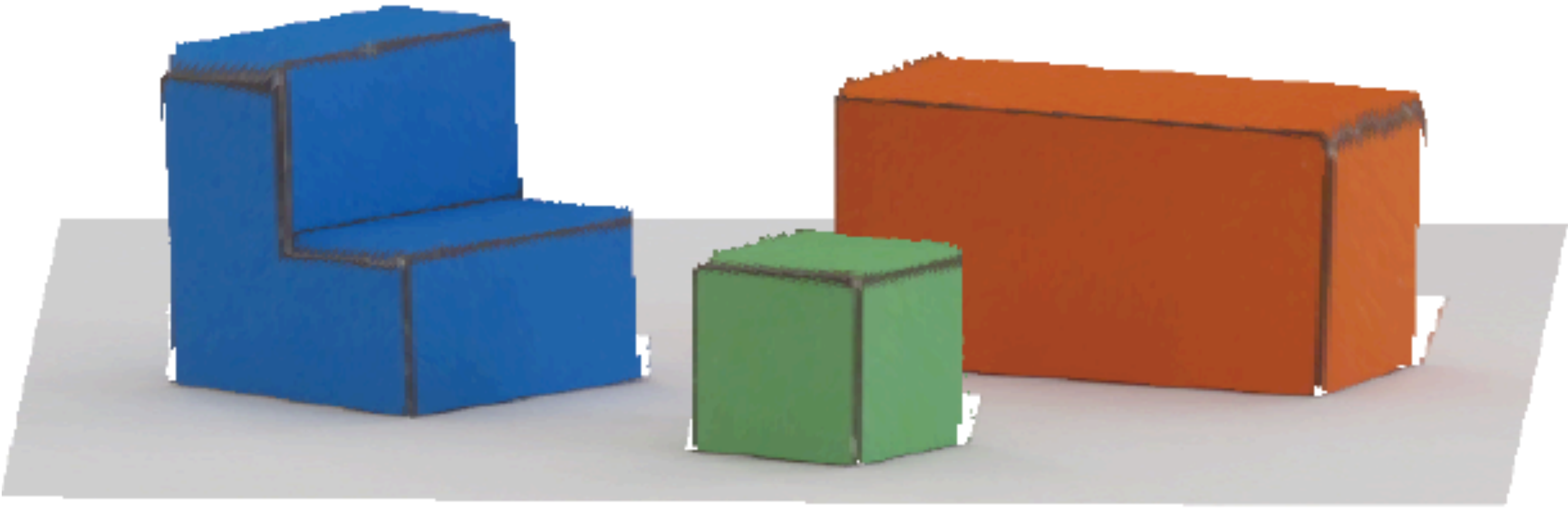
Changing view point

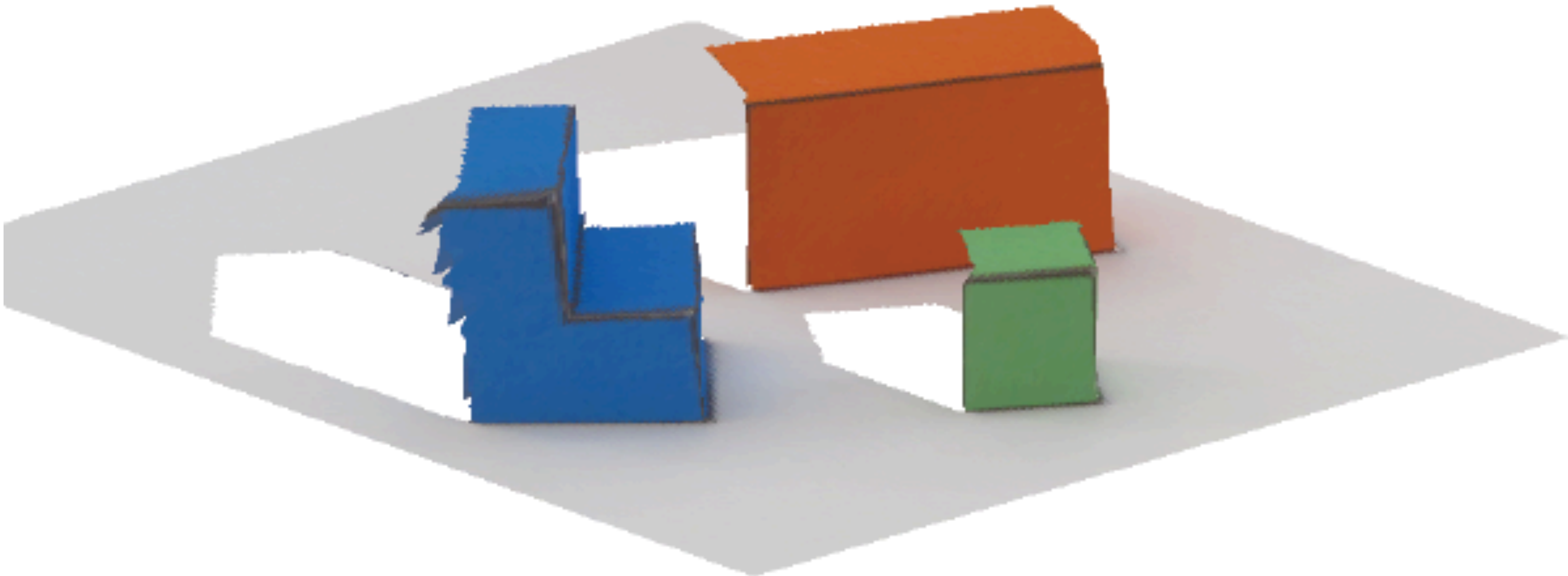
Input



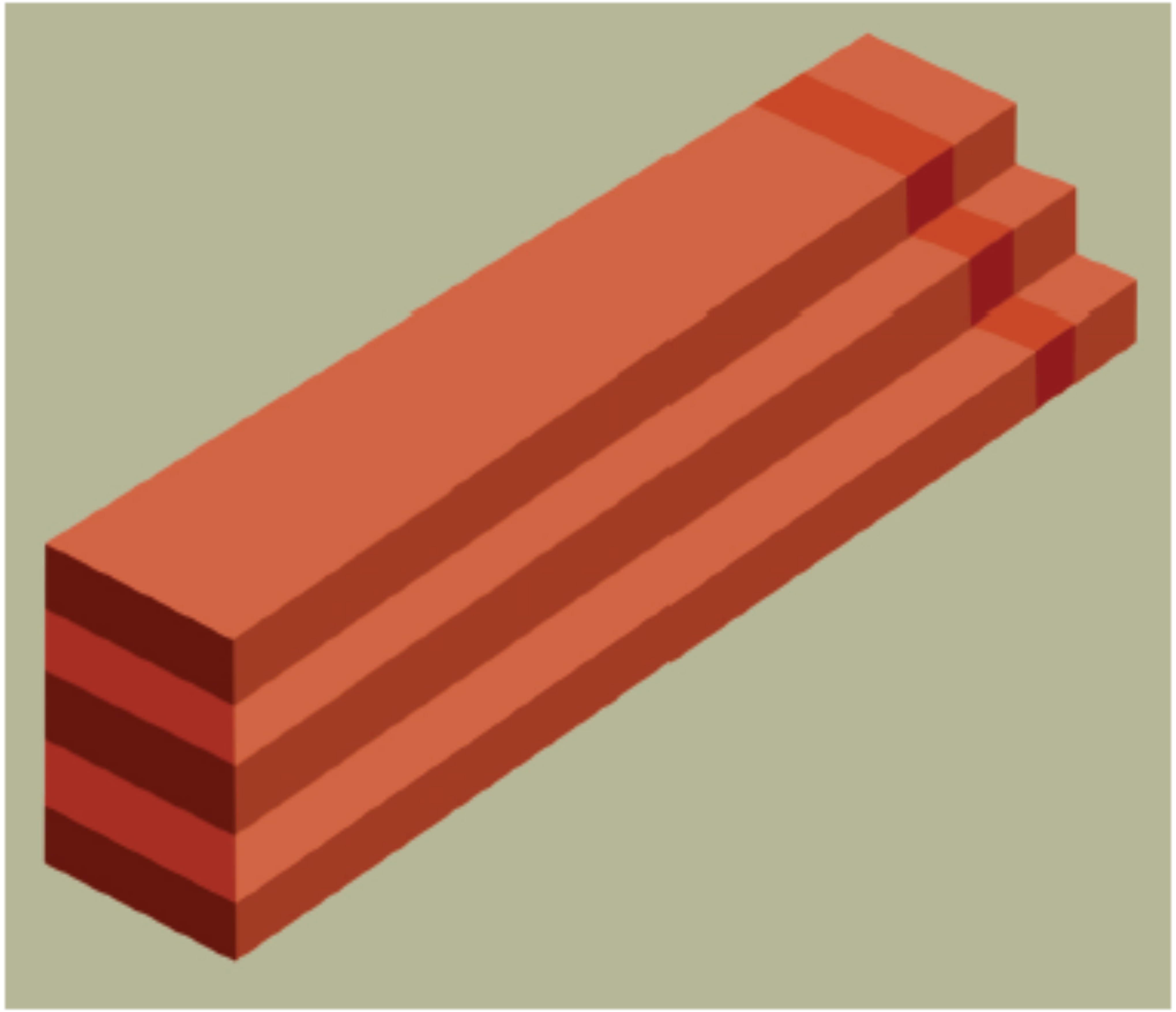
New view points:

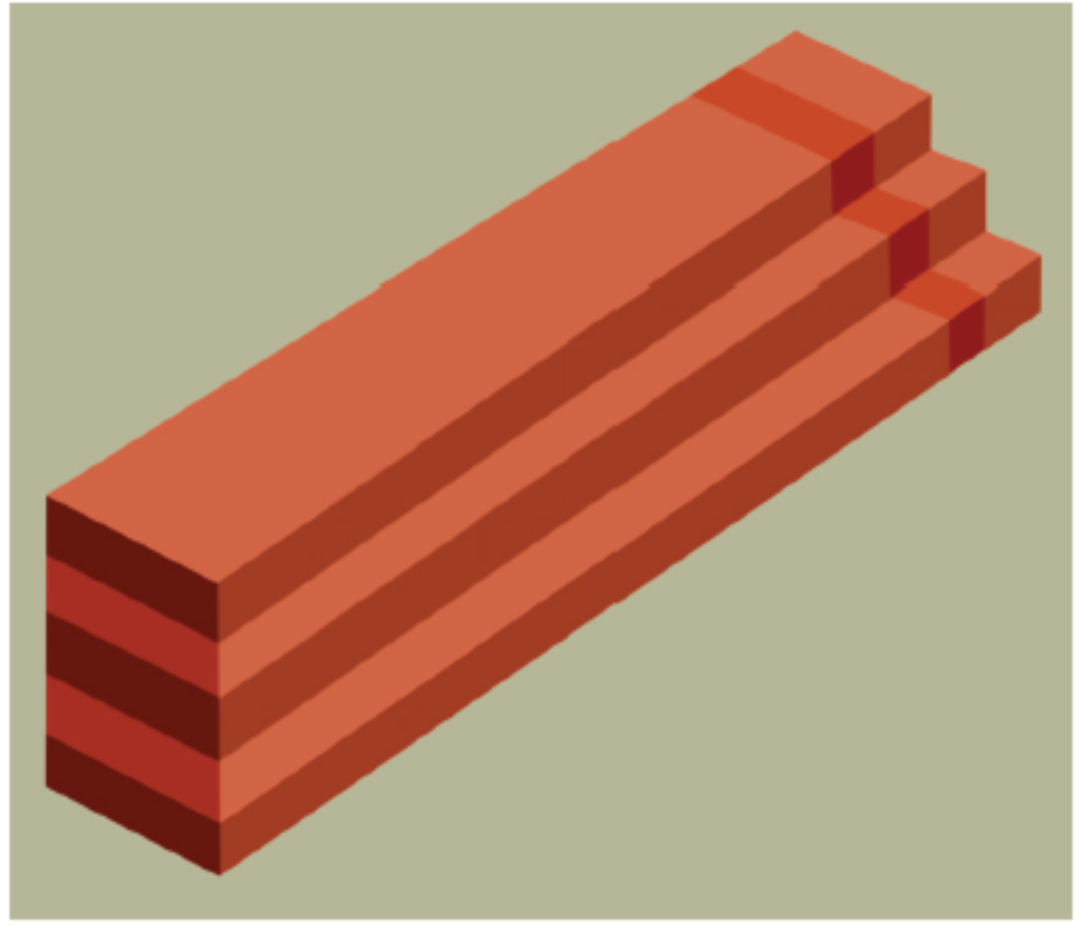




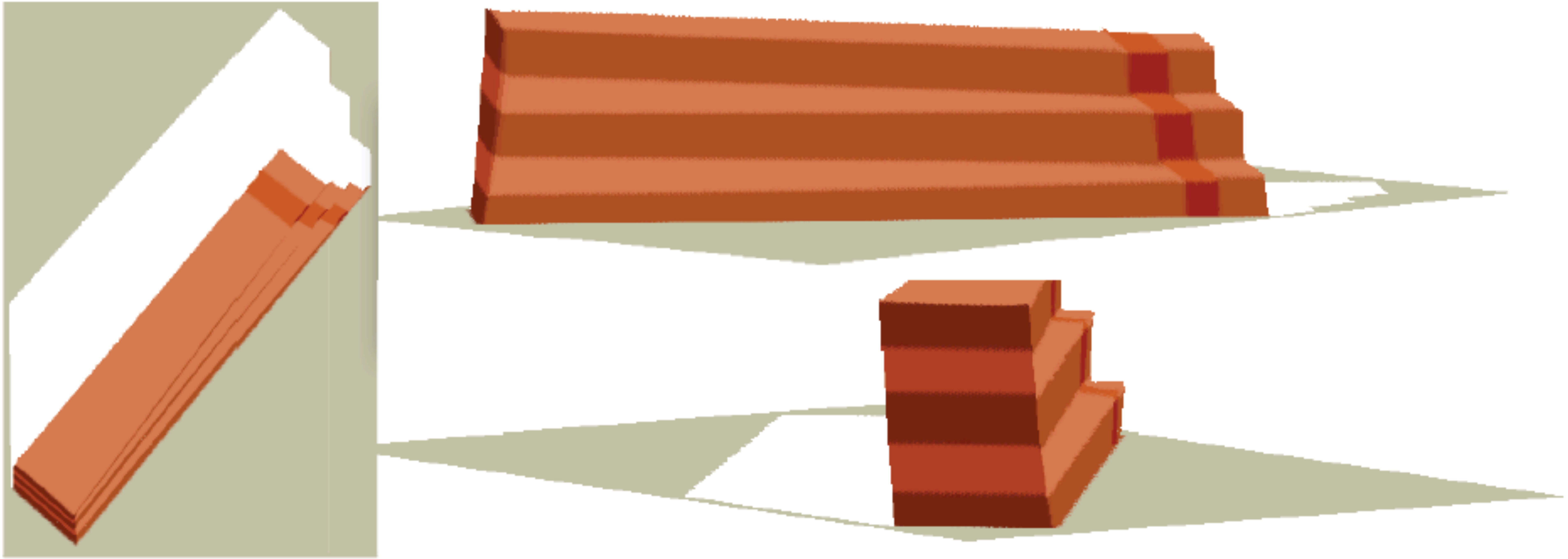


Impossible steps



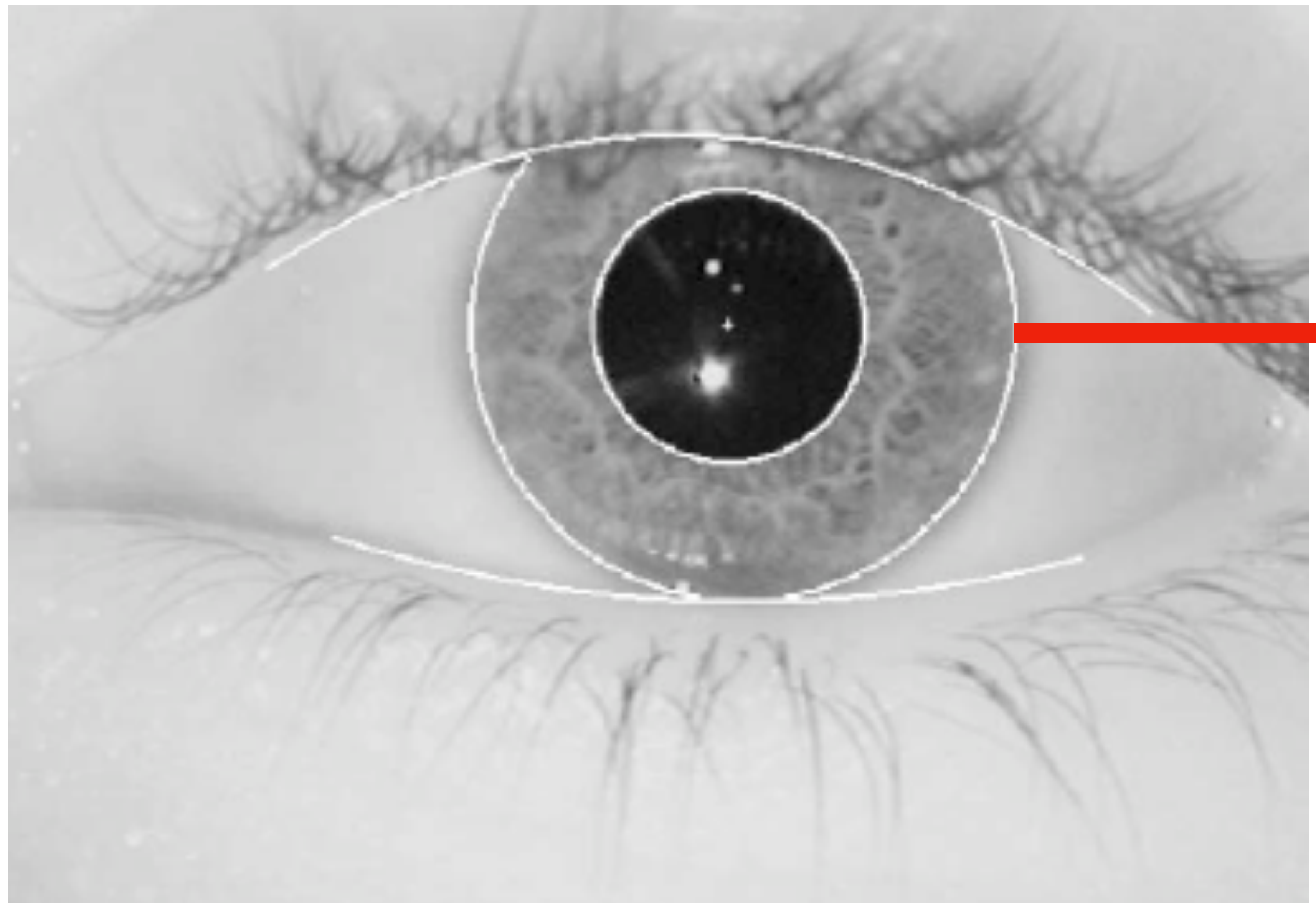


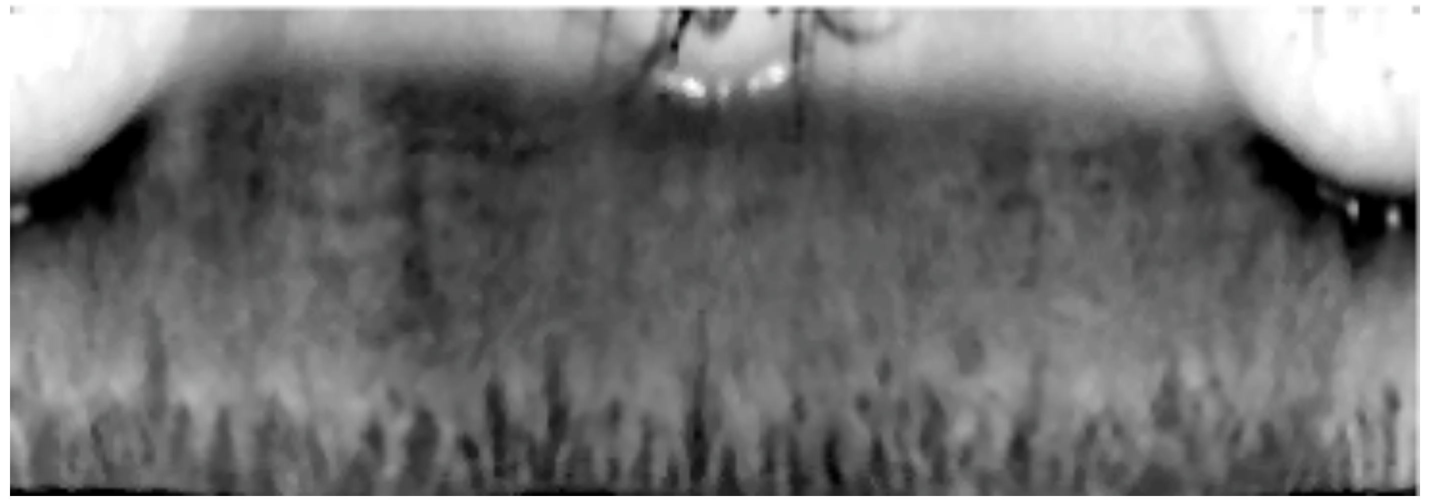
Impossible steps



Tasks: what machines are good at

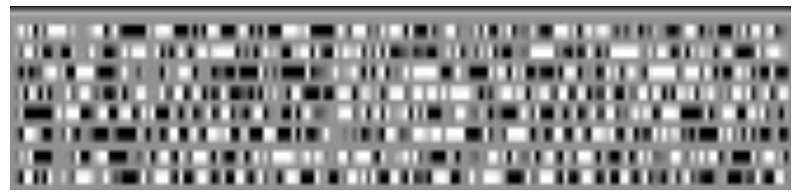
Seeing small and precise details

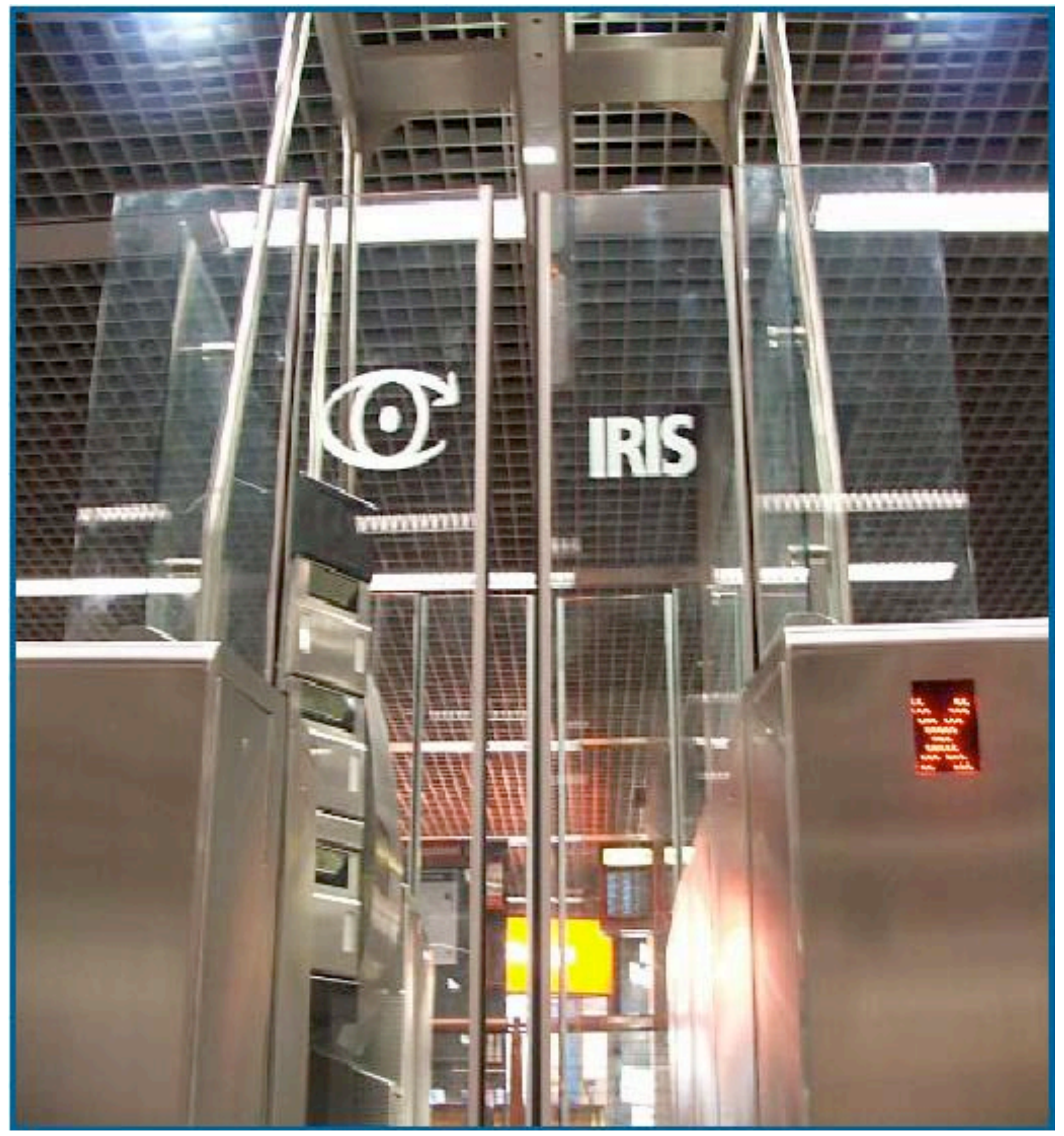




Unwarped Iris image

Iris code

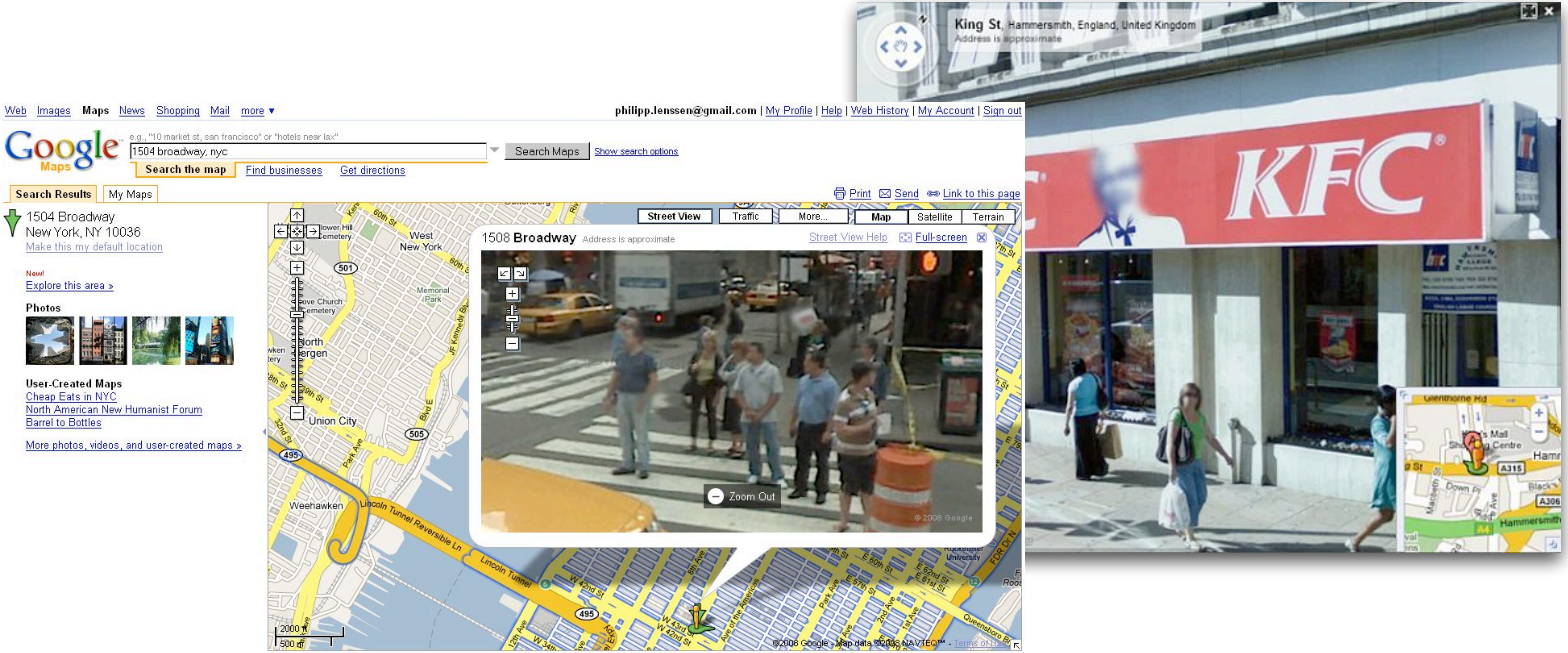




http://www.ind.homeoffice.gov.uk/managingborders/technology/iris/

Tasks: what machines are good at

Doing many times the same thing without losing attention



Tasks: what machines are good at

Tasks: what machines are good at

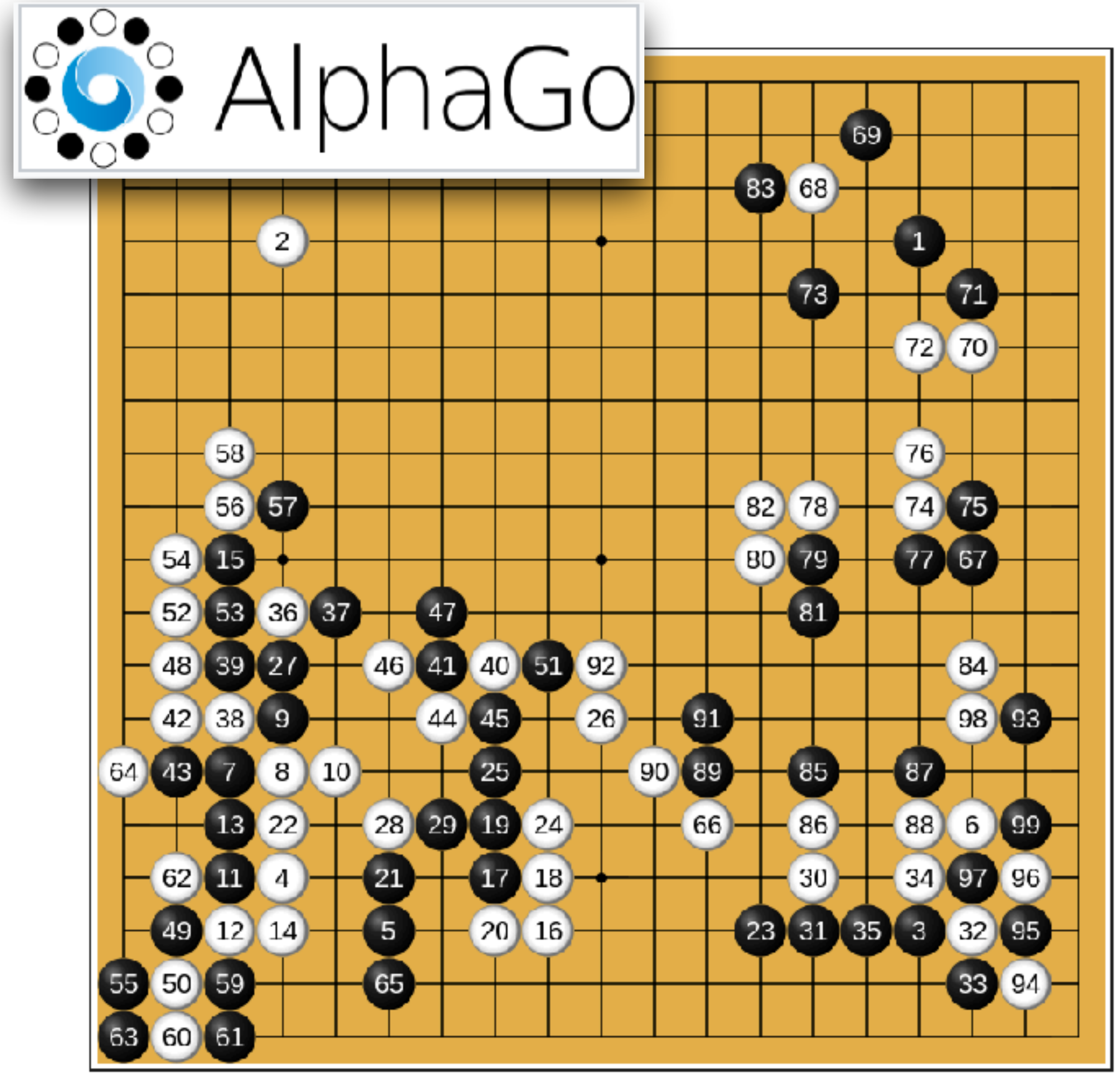
(a) Image 1

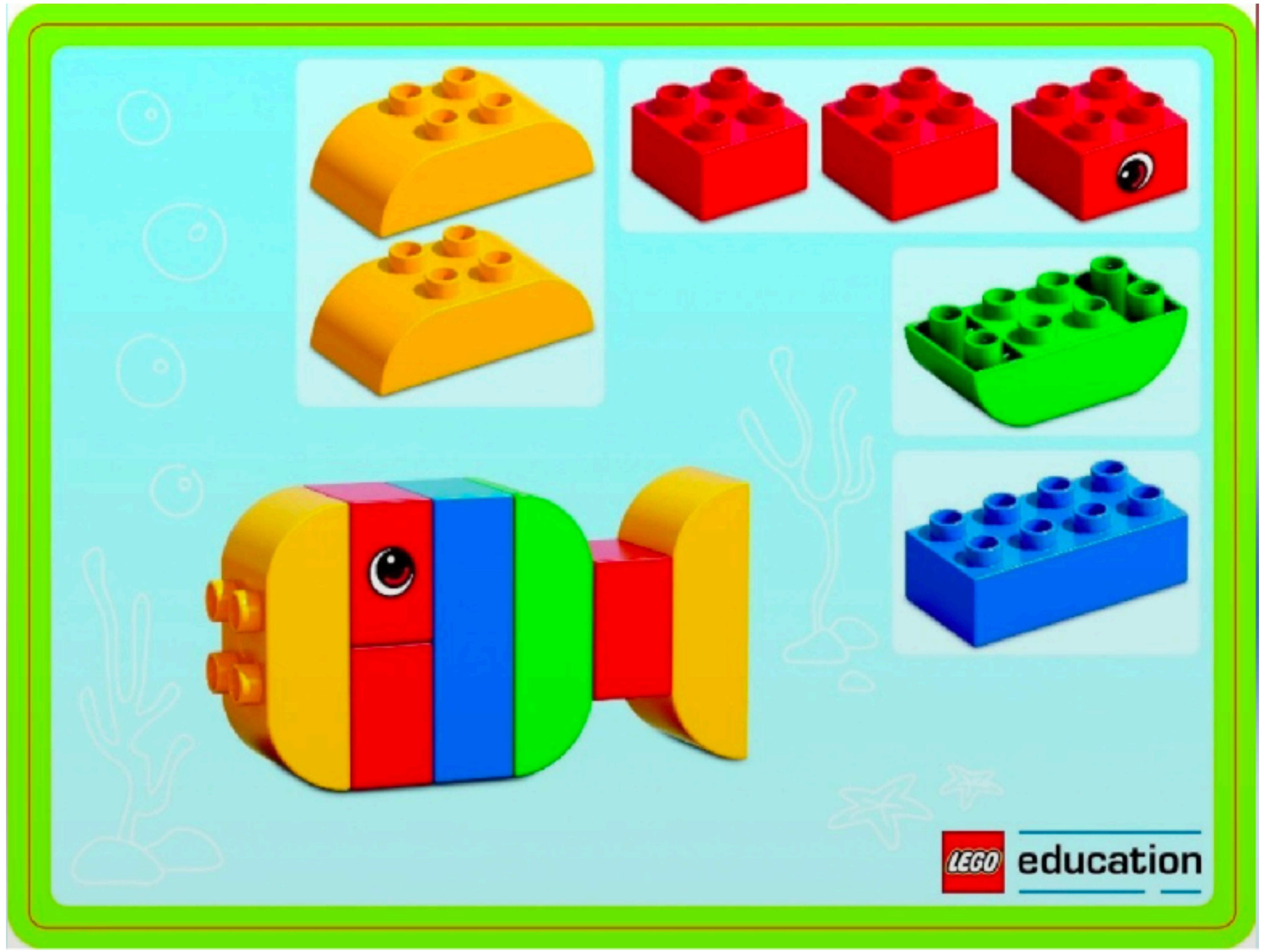
(b) Image 2

Image stitching

Tasks: what machines are not so good at

Great at tasks that require exact computations, memory and exploration

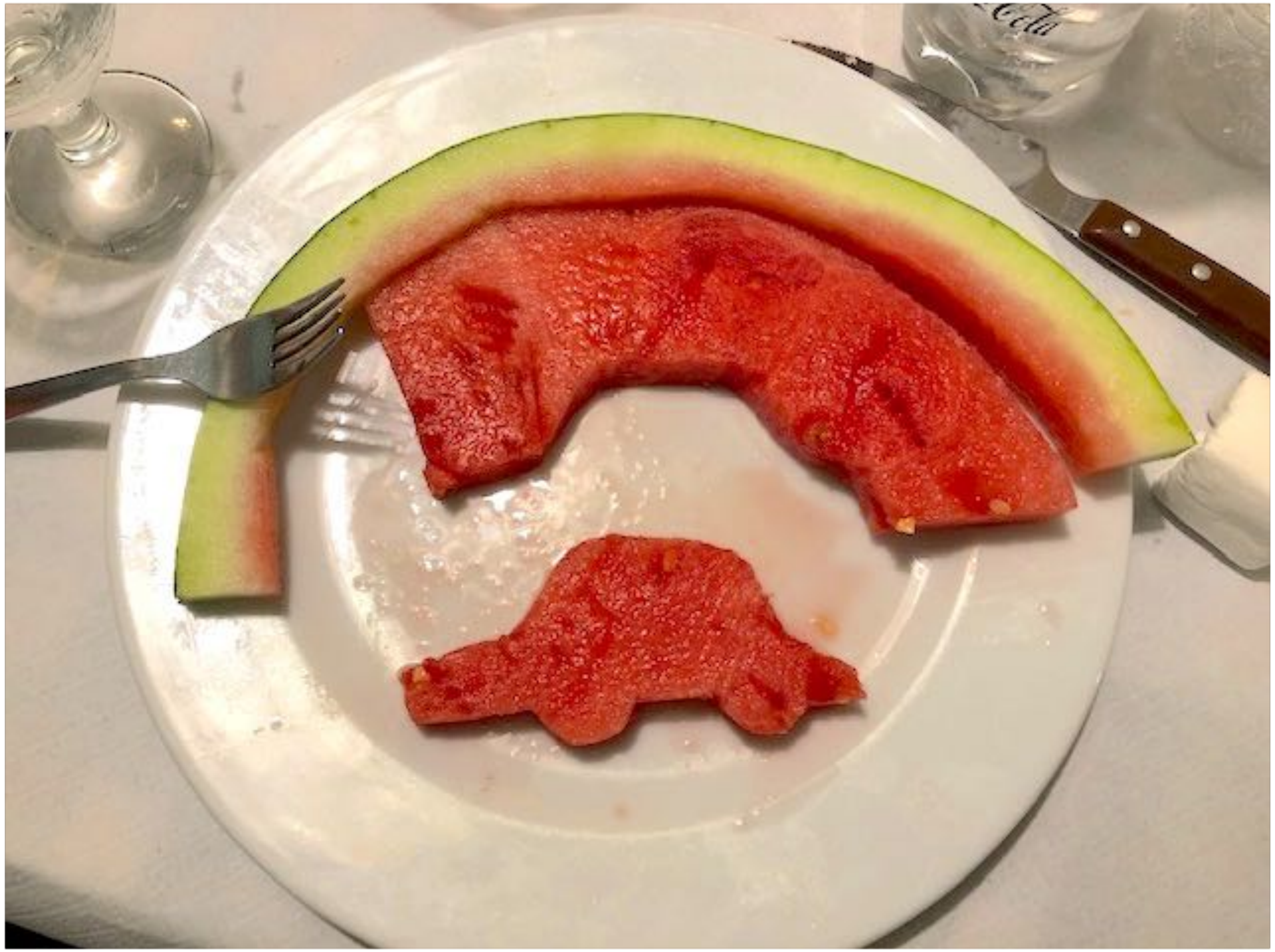




Not so good on common sense reasoning, like what is needed for going from an arbitrary set of visual instructions to behavior.

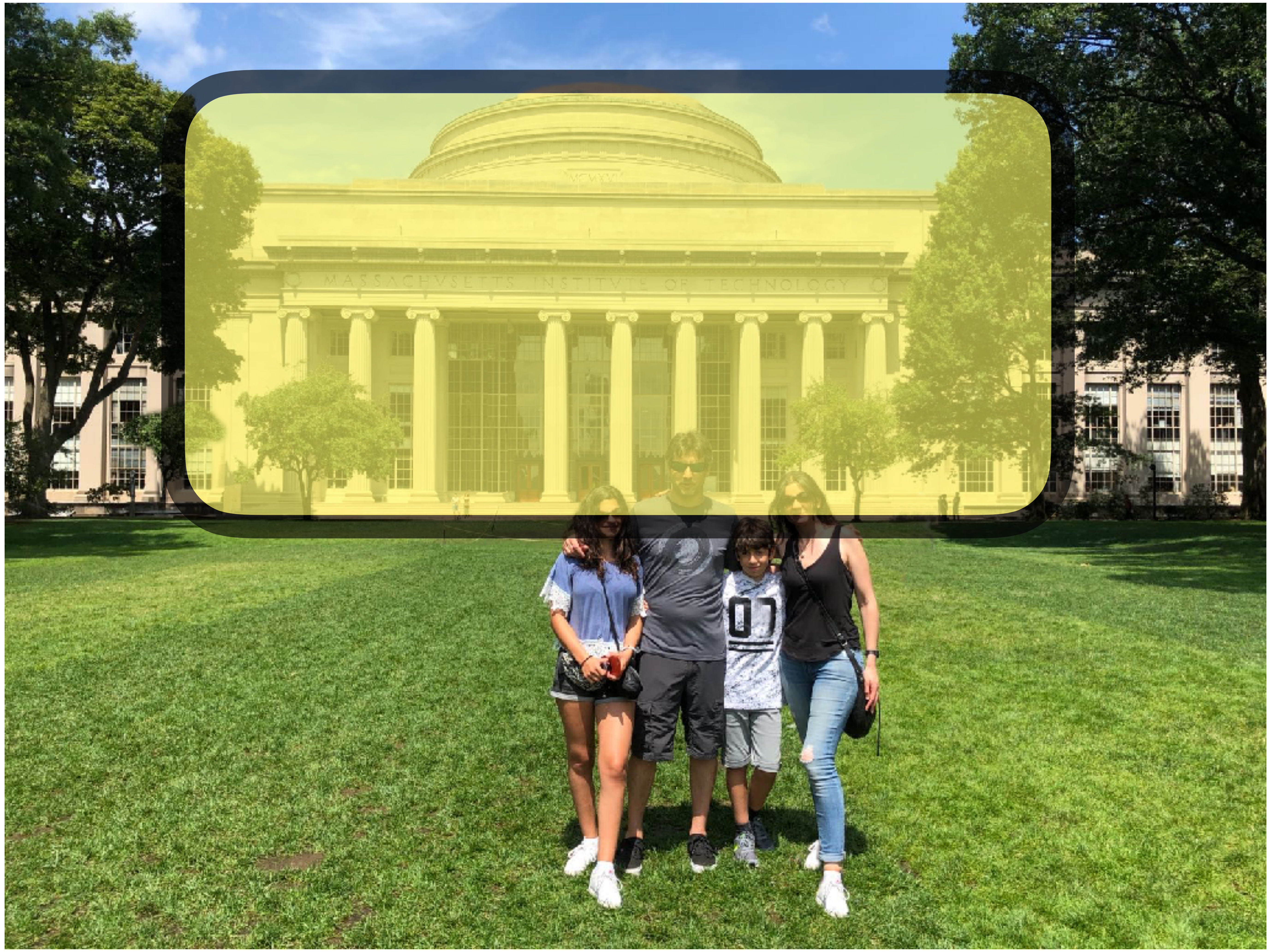
Tasks: what machines are not so good at

Visual intelligence, reasoning



Tasks: what humans care about

Tasks: what humans care about



Verification: is this a building?

Recognition: which building is this?

Tasks: what humans care about

Image classification: list all the objects present in the image

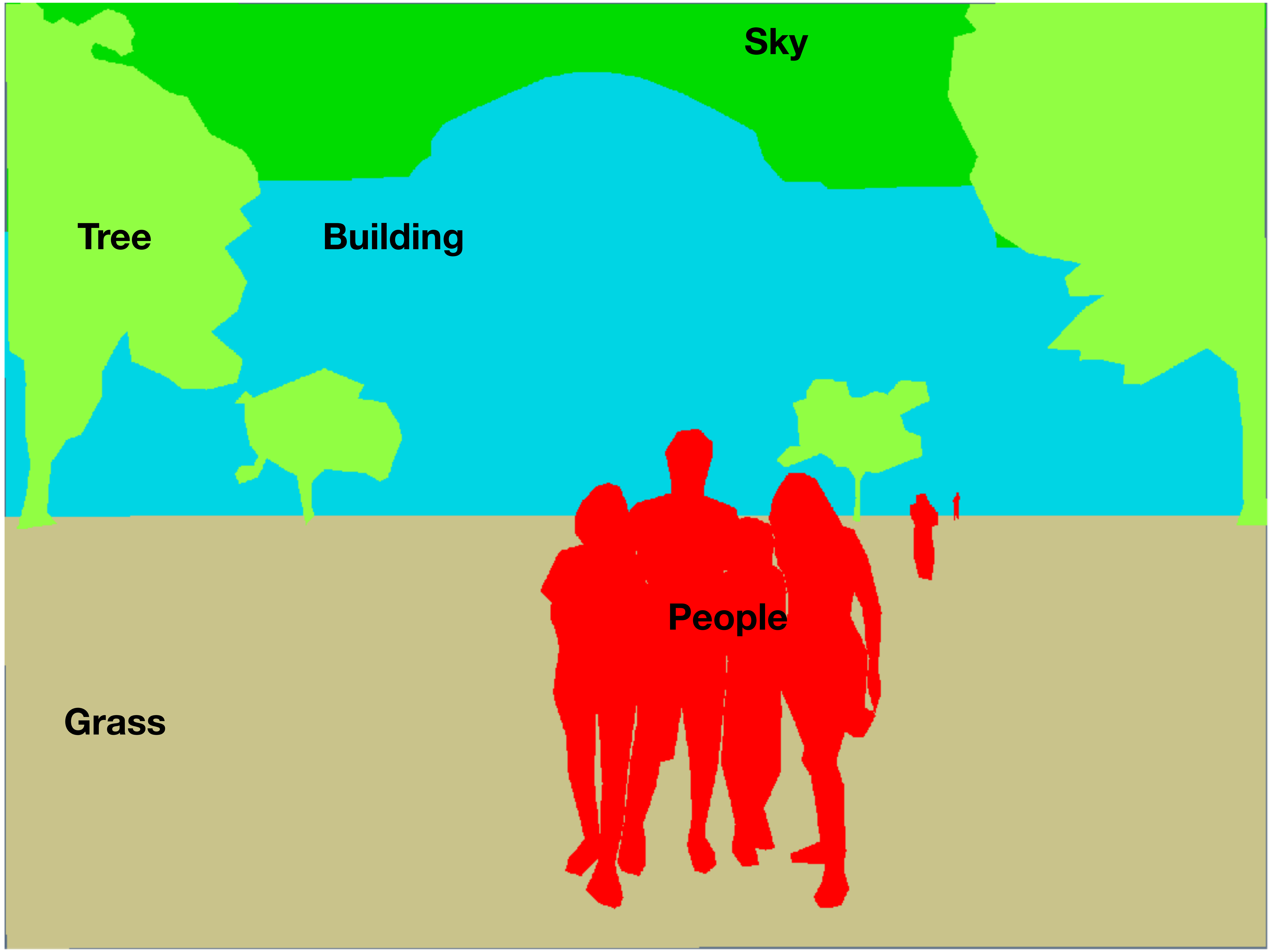
- Building
- Grass
- People
- Trees
- Sky
- Columns
- …

Tasks: what humans care about

Scene categorization

- Outdoor
- Campus
- Garden
- Clear sky
- Spring
- Group picture
- …

Tasks: what humans care about

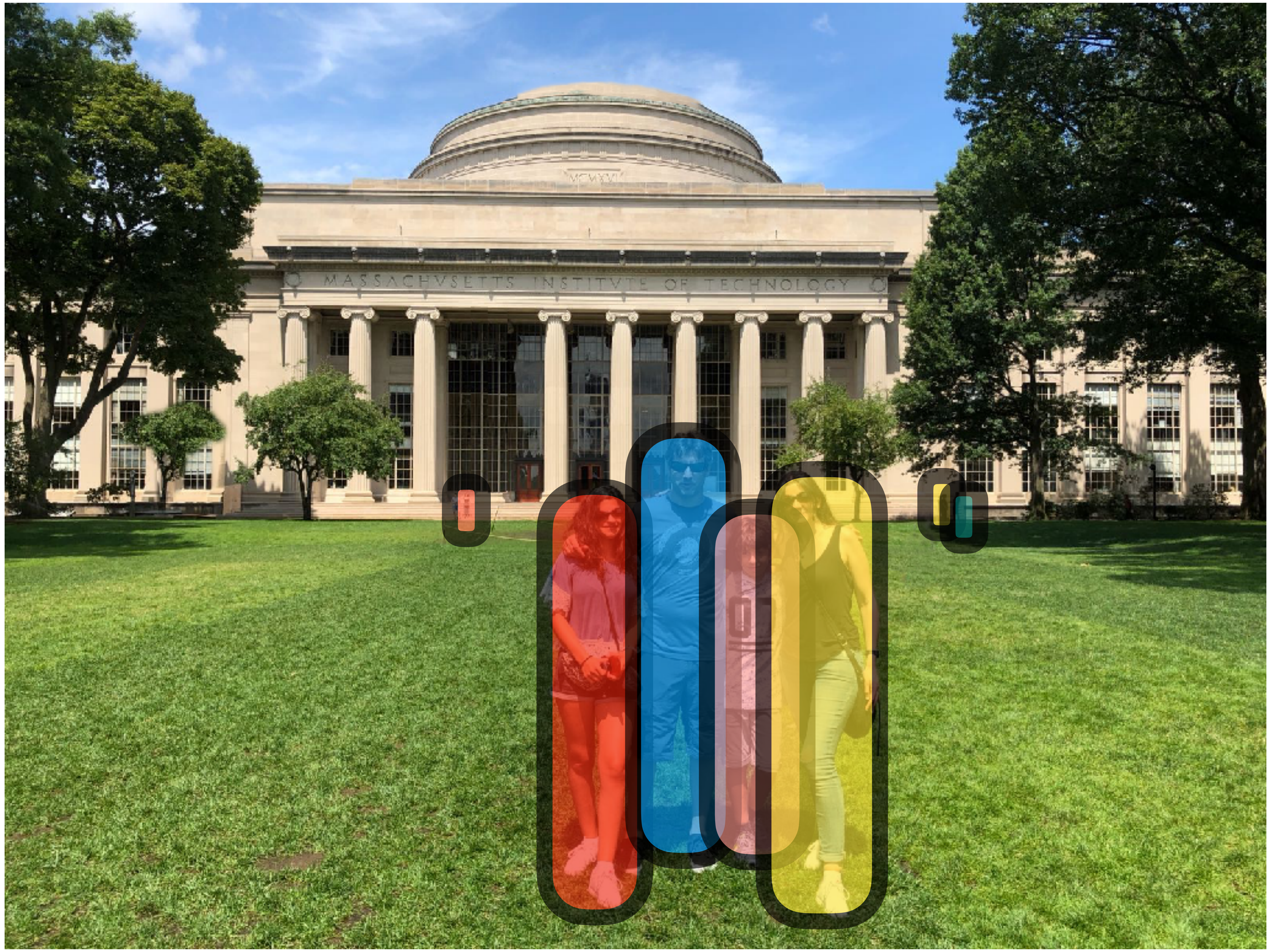


Semantic segmentation:
Assign labels to all the pixels in the image

Related tasks:

- **Semantic segmentation**
- **Object categorization**

Tasks: what humans care about

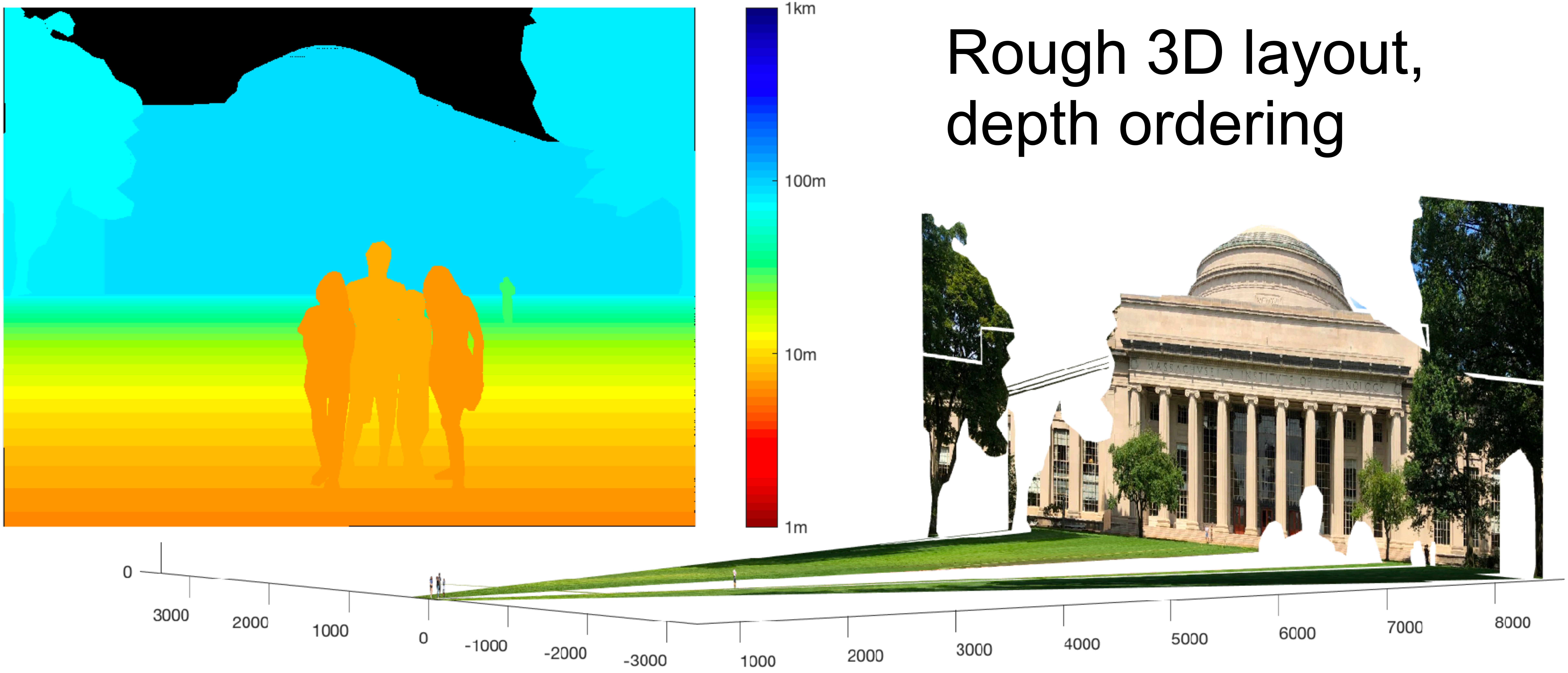


Detection: Locate all the people in this image

Tasks: what humans care about

Recognition: who is this person?

Tasks: what humans care about



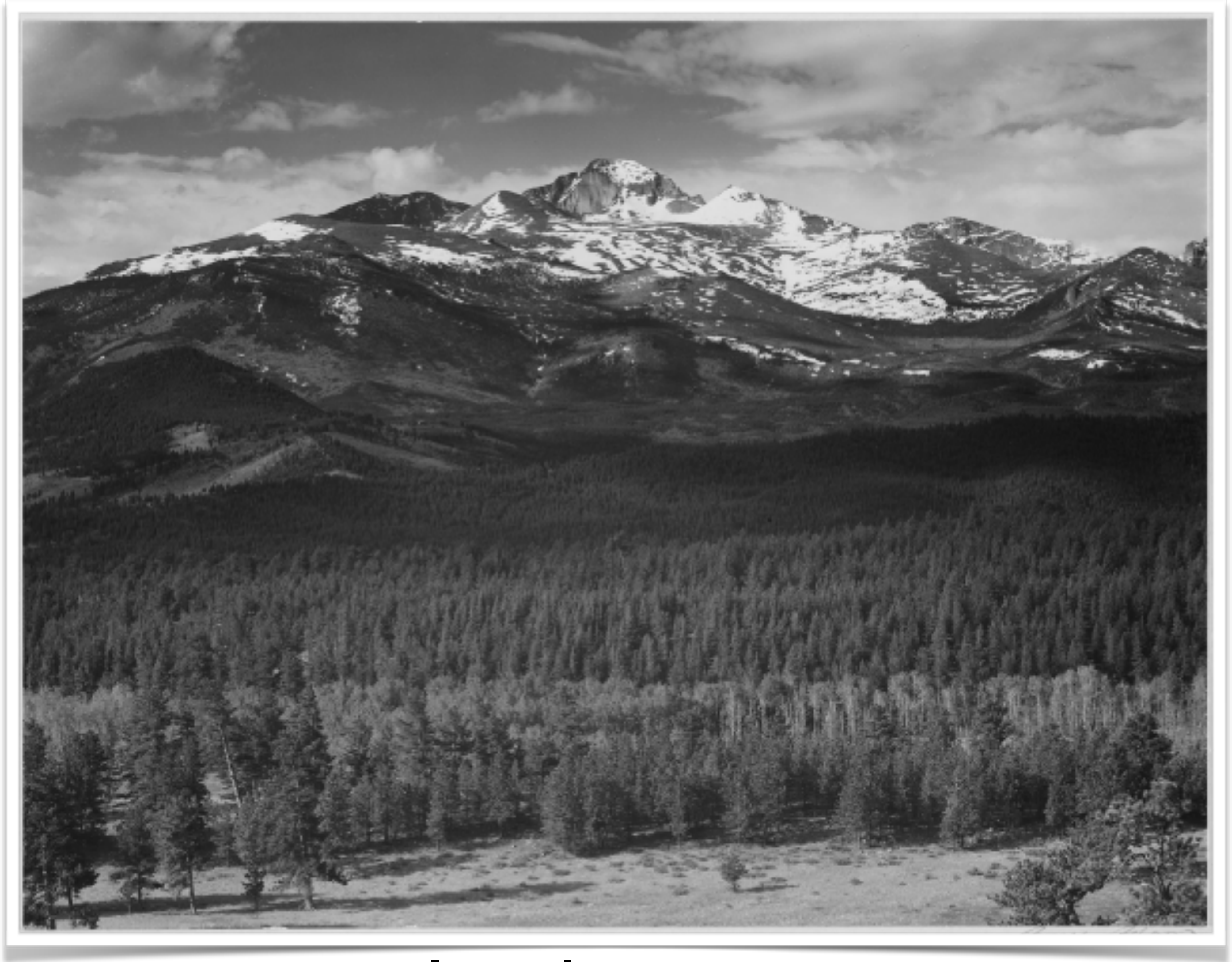
Rough 3D layout, depth ordering

Tasks: what humans care about

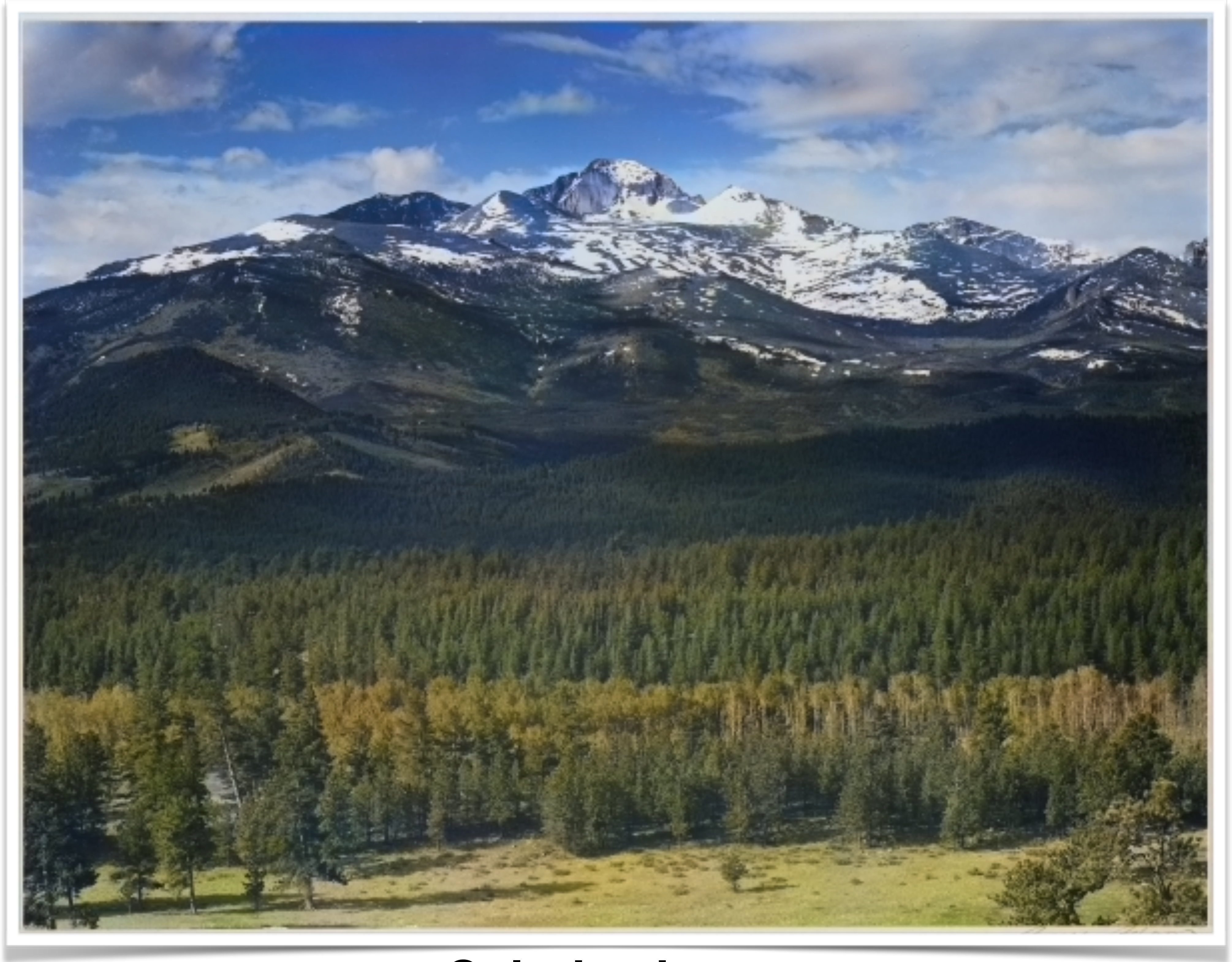
Making new images

Tasks: what humans care about

Adding missing content



Input image



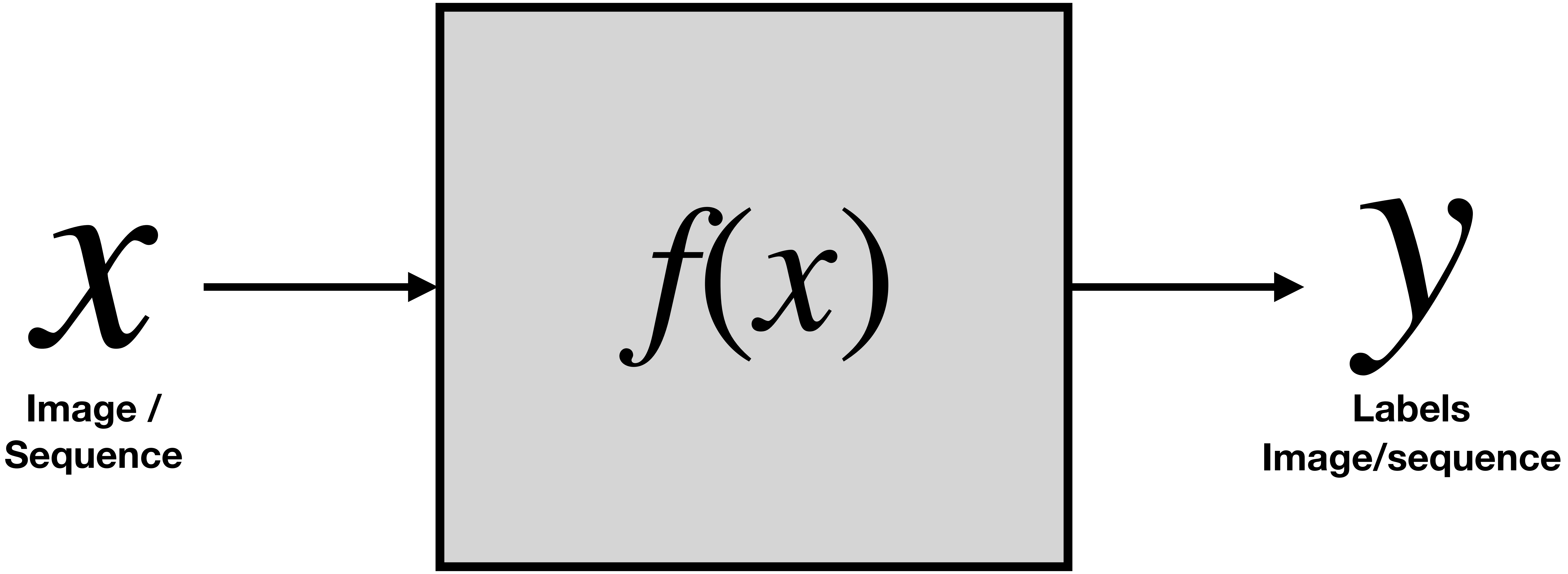
Colorized output

Tasks: what humans care about

Predicting future events

What is going to happen?

Tasks: generic formulation



1. Introduction to computer vision

- History
- Perception versus measurement
- Simple vision system
- Taxonomy of computer vision tasks