Computer Vision: The Quick Tour

Anoop M. Namboodiri
CVIT, IIIT Hyderabad

Outline

- Introduction to Computer Vision [20 minutes]
  - What, why, why not?
- Camera Model and Geometry [20 minutes]
- Problems in Computer Vision
  - Recovering world geometry [20 minutes]
  - Reorganizing images [20 minutes]
  - Detection and Recognition [20 minutes]
- Questions and Discussions [10 minutes]

What is Computer Vision?

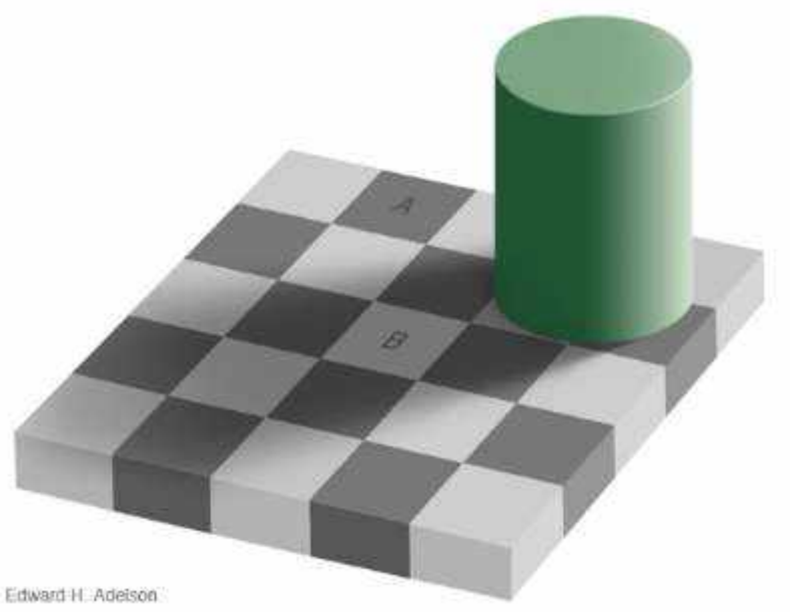
- Understanding of visual inputs (images/videos) by computers.
- Making sense out of them. Describing them.
- Does computer vision mimic the human vision?
  - Certainly in many of its goals
  - Why? Human vision is among the best!
  - Sophisticated and efficient but not understood well
- Should computers process visual inputs like humans?

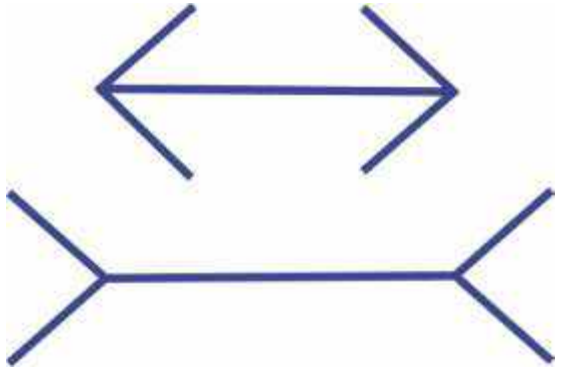
Not necessarily!

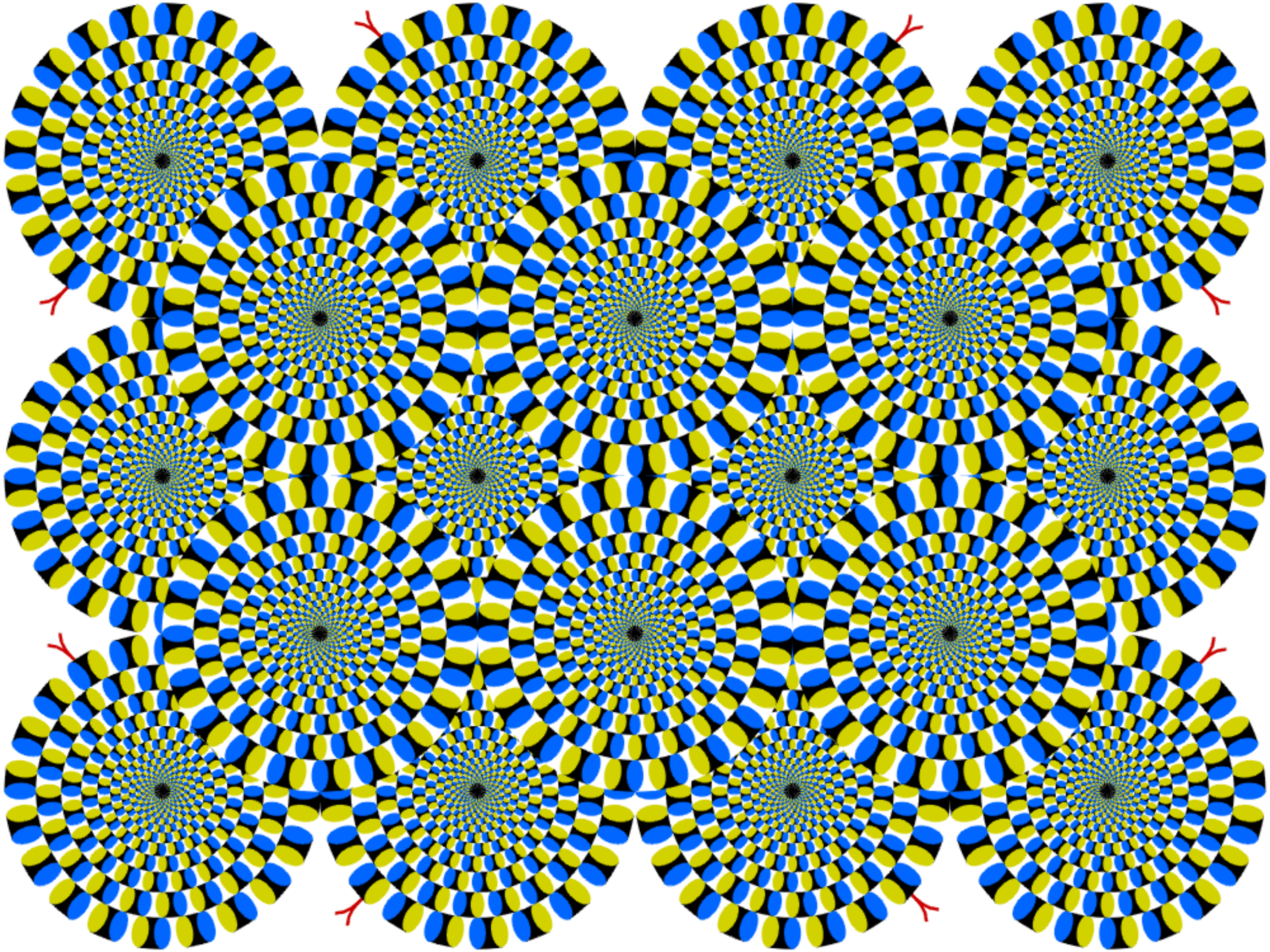
  - Human visual system need not limit computer vision
  - We draw inspiration from it as often as is convenient

Human perception has its shortcomings…

Sinha and Poggio (Image: Ron Rensick)







Three "Urges" on seeing a Picture*

1. **To group** proximate and similar parts of the image into meaningful "regions".

   Called segmentation in computer vision.

2. **To connect to memory** to recollect previously seen "objects".

   Called recognition in computer vision.

3. **To measure** quantitative aspects such as number and sizes of objects, distances to/between them, etc.

   Called reconstruction in computer vision.

**Jitendra Malik; Mysore Park, Dec. 2011*

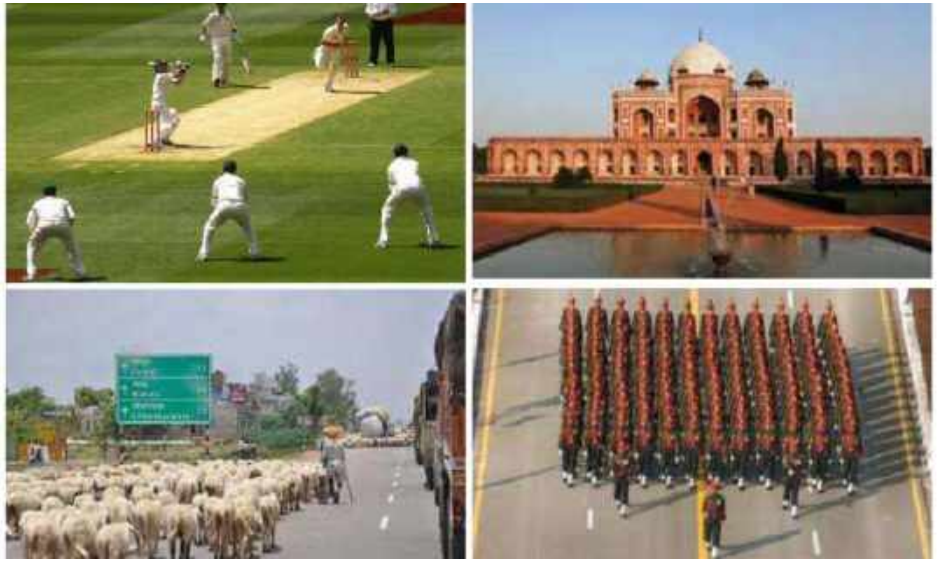
The Three Rs of Computer Vision

Reorganization (Segm.)

Group semantically similar pixels

Recognition

Connecting what we see to our memory



Reconstruction

Measure/recreate a 3D model of what we see in the world

Why is it Difficult?

90	126	180	120	102	131	126	91
82	140	143	182	180	142	138	81
81	141	148	195	188	147	140	80
75	144	150	210	198	149	141	73
71	144	151	241	214	150	143	70
88	142	147	236	205	146	141	85
106	139	142	225	197	141	138	101
128	135	139	184	180	138	132	121

Computer Vision

- Goal: Extract all possible information about a visual scene by computer processing

  What? When? Where? Who? How? Why? How many?

- Over 50% of the brain is devoted to vision for humans.
  – Must be important to us!

- Why is it difficult?

Chairs and Chairs

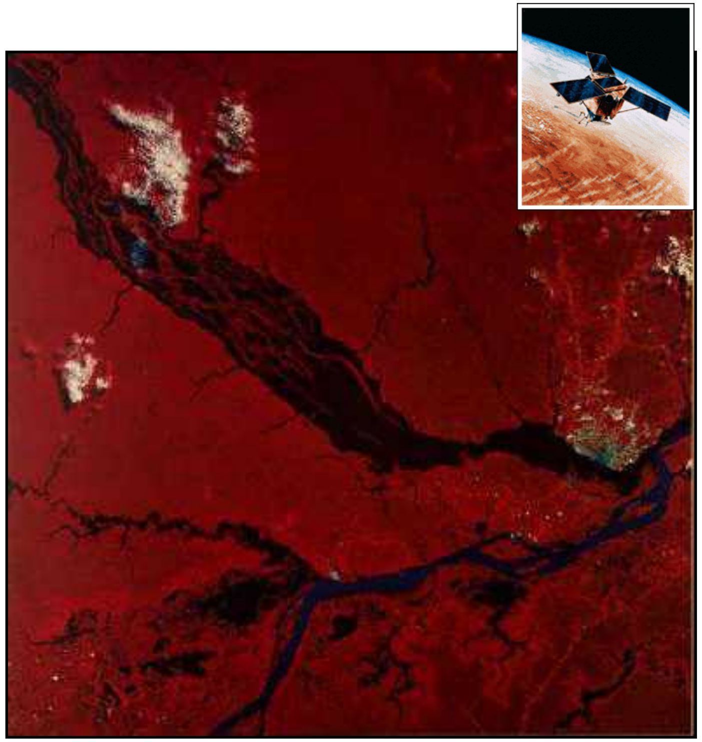
- Which are chairs?
- Large intra-class variations
- How do we describe a chair?
- Basic property: Sittability!
- We infer a lot from pictures. Can we instruct a computer to do the same?
- Do we understand how we infer?

Why Automated Vision?

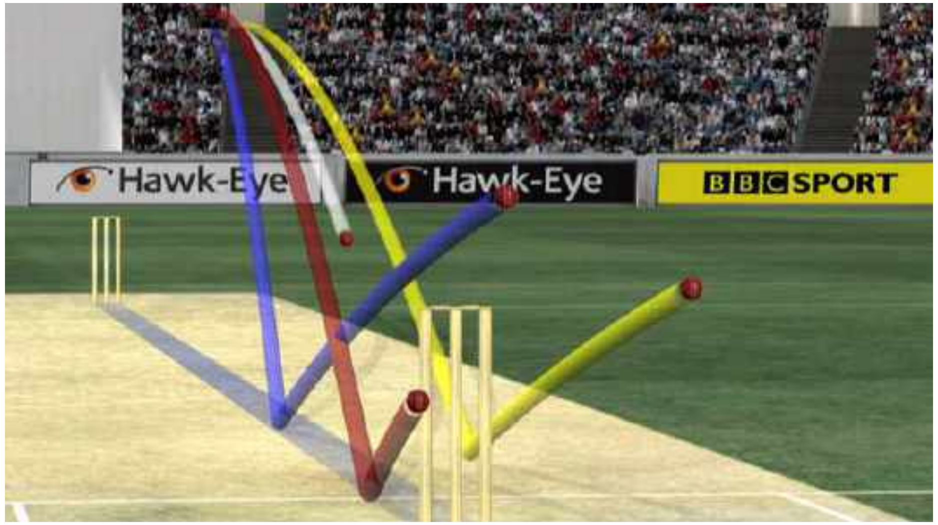
1. High reliability
2. High repeatability
3. More objective evaluation
4. Lower cost
5. Higher speed
6. Ability to operate in hazardous environments

General purpose machine vision system do not exist.

Applications

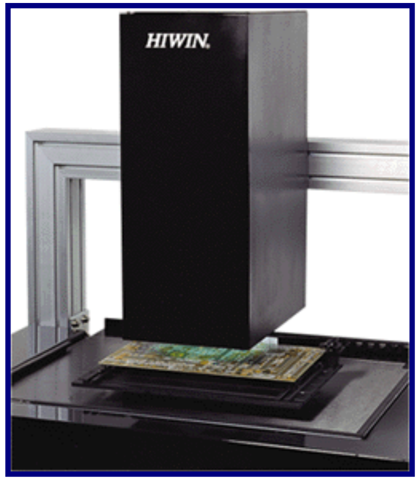


Rio Negro (black), Amazon (blue)



Ball Tracking: Hawk Eye

Segmentation



Manual vs. Automated PCB Inspection

And many many more..

- Surveillance
- Automated Assembly
- Mail Sorting
- Face detection (photography)
- Robot Navigation
- Content-Based Info. Retrieval
- Movies
- Logistics
- Traffic control
- Automotive Safety
- Medical Diagnostics
- Building Automation
- Gaming
- Broadcasting (infographics)
- Crowd Control
- Agriculture
- …

and last bust not the least..

Autonomous Navigating Cars

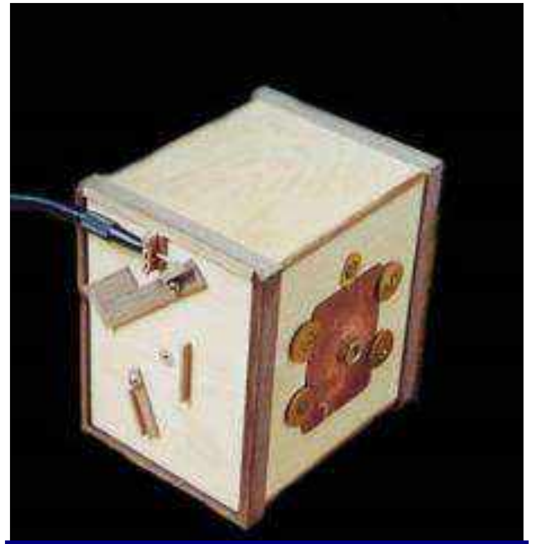
The Real Problem



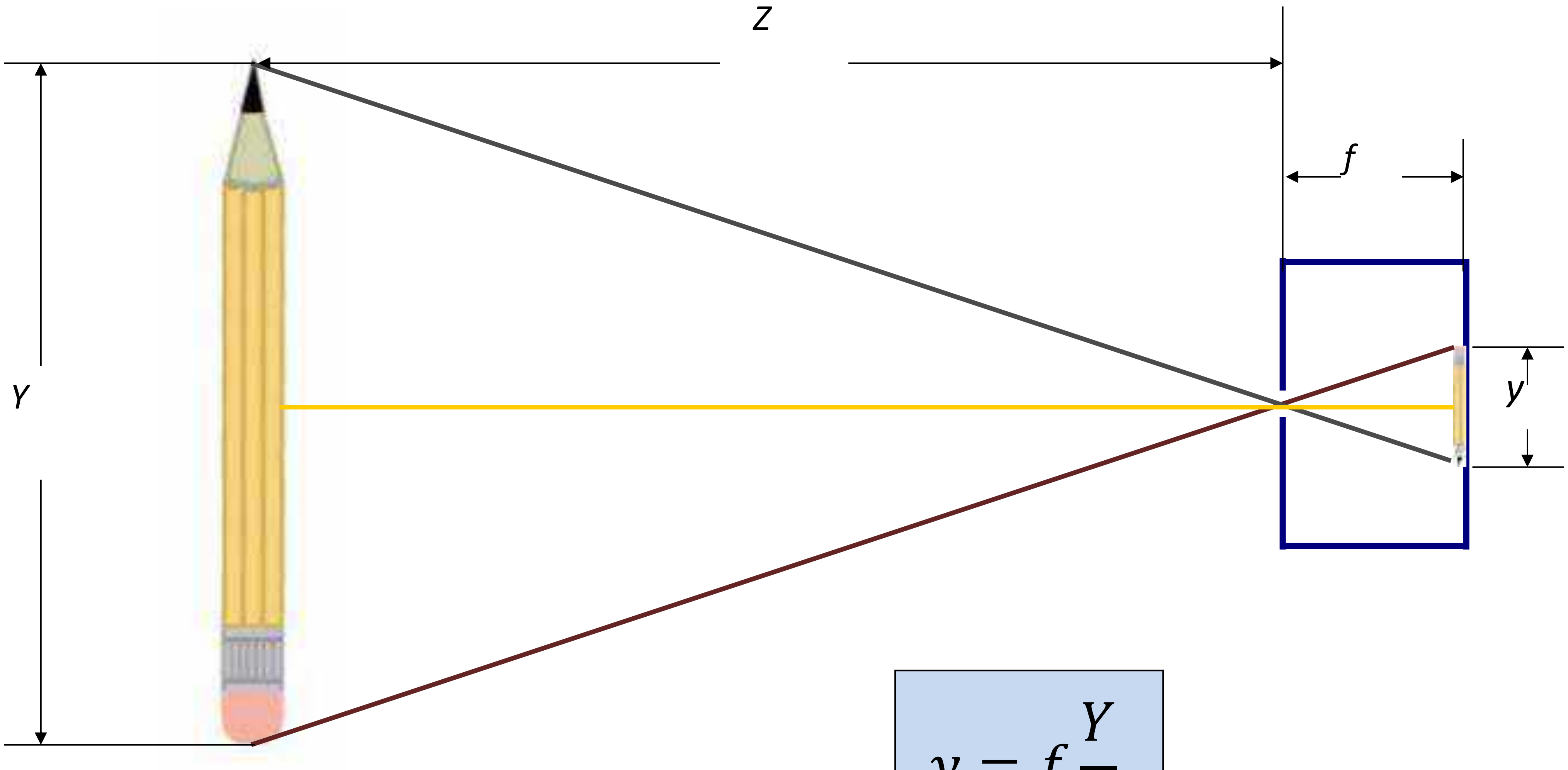



Outline

- Introduction to Computer Vision [20 minutes]
  - What, why, why not?
- Camera Model and Geometry [20 minutes]
- Problems in Computer Vision
  - Recovering world geometry [20 minutes]
  - Reorganizing images [20 minutes]
  - Detection and Recognition [20 minutes]
- Questions and Discussions [10 minutes]

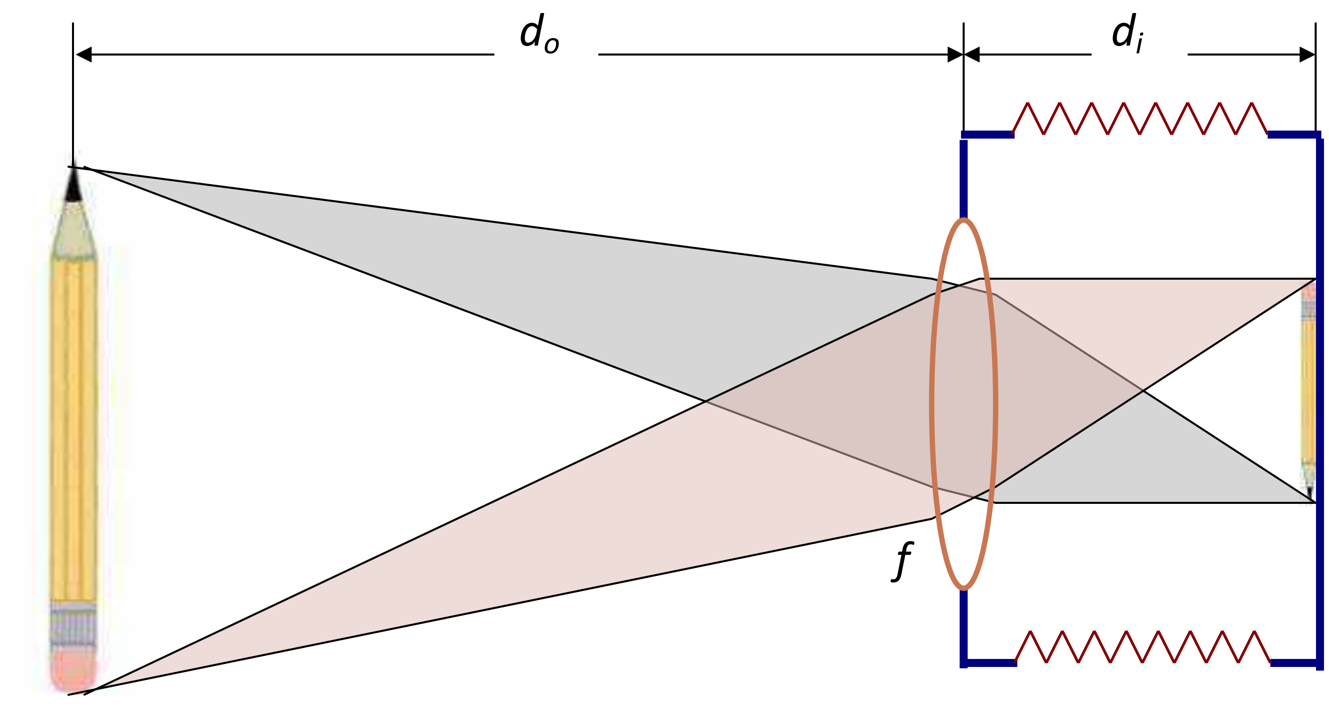


The Pinhole Camera



$$y = f\frac{Y}{Z}$$

Camera with Lens

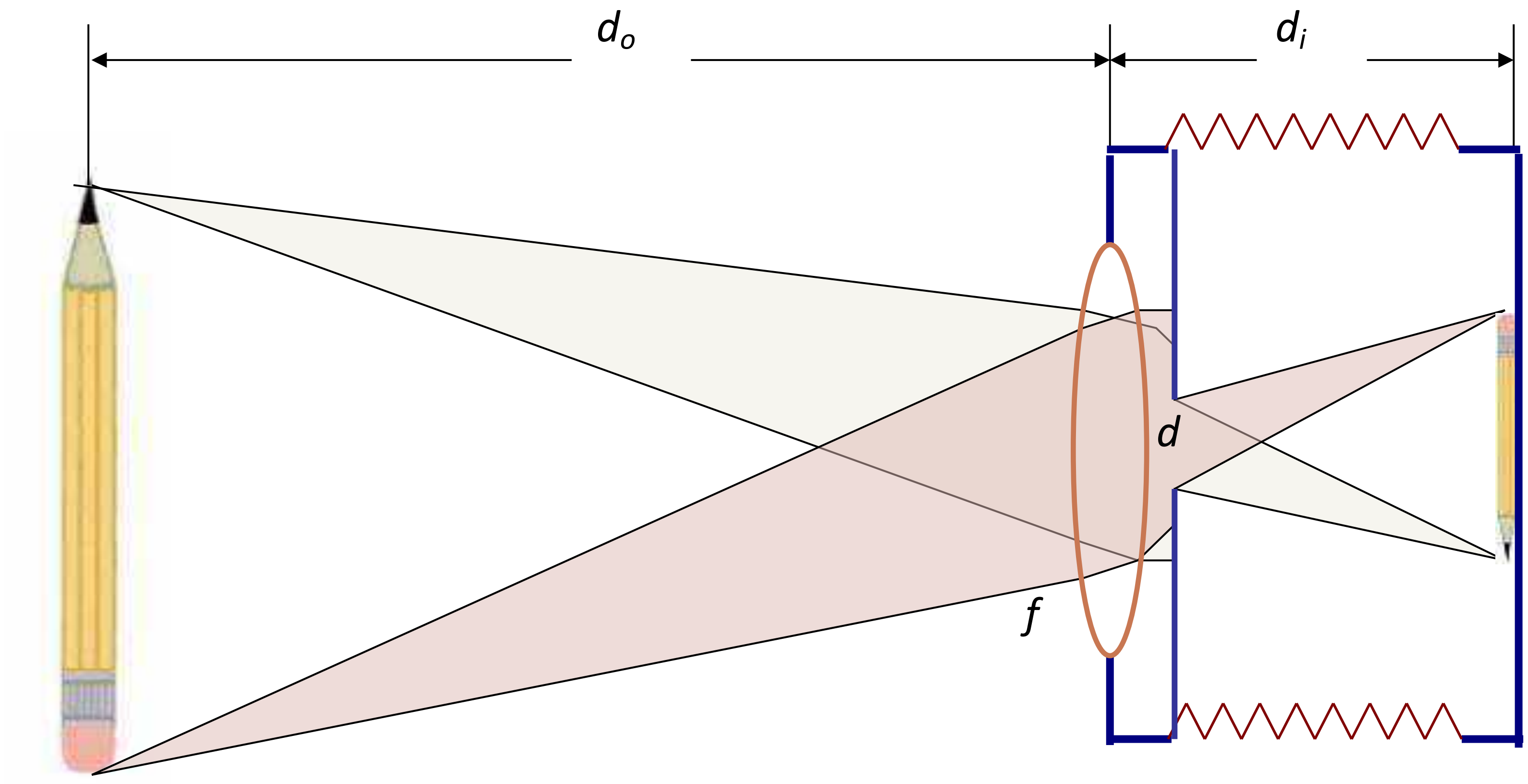


Thin lens equation: $\dfrac{1}{f} = \dfrac{1}{d_o} + \dfrac{1}{d_i}$

$$d_i = f\,\frac{d_o}{(d_o - f)}$$

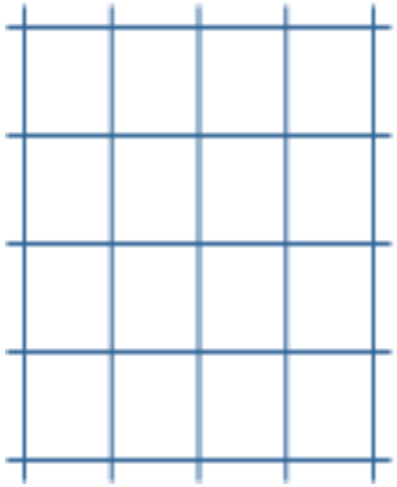
Focus and DOF

Aperture

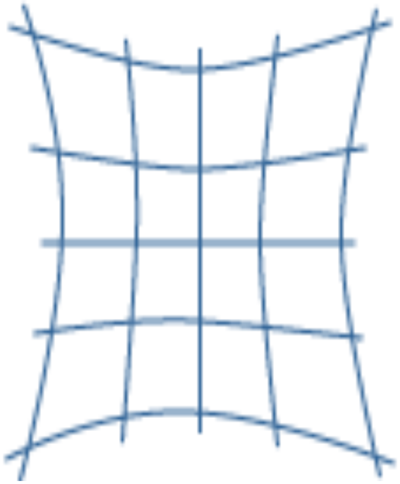


Focal Ratio = f / d

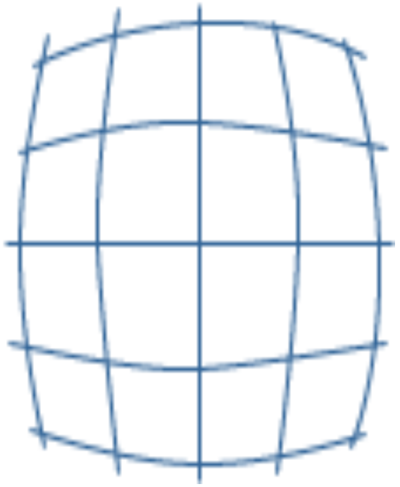
Geometric Distortions



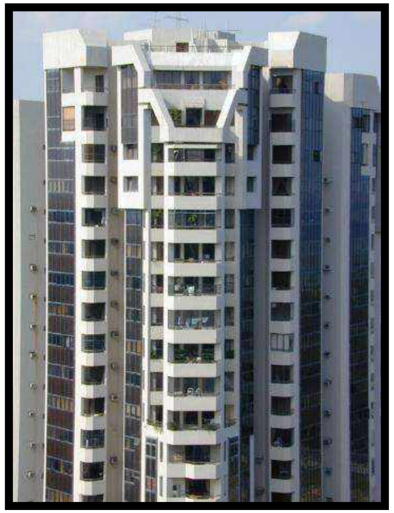
Ideal



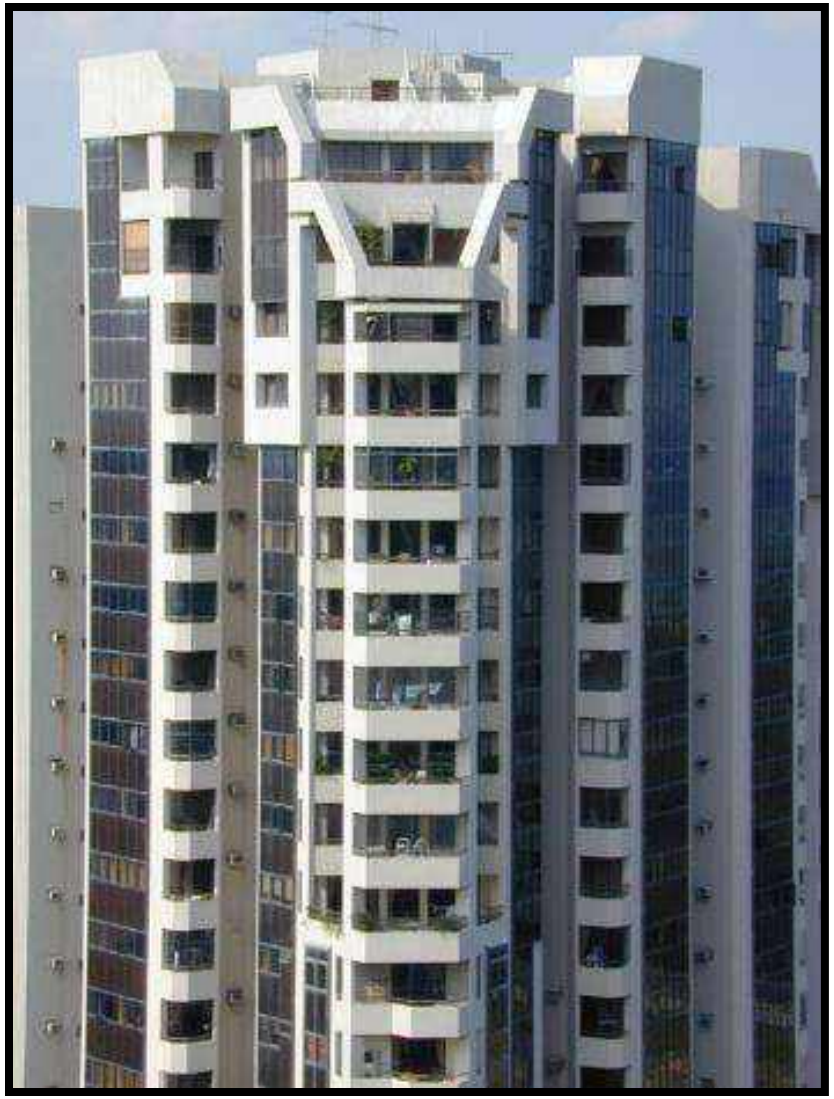
Pincushion



Barrel

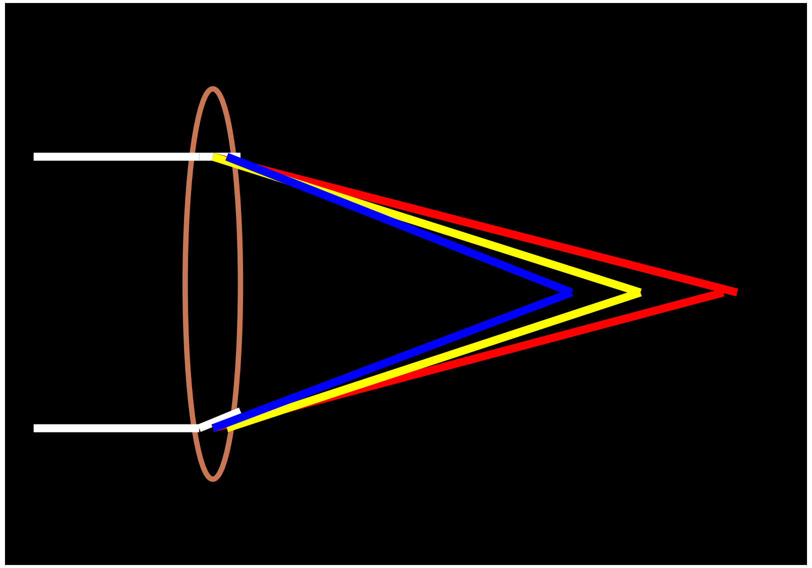


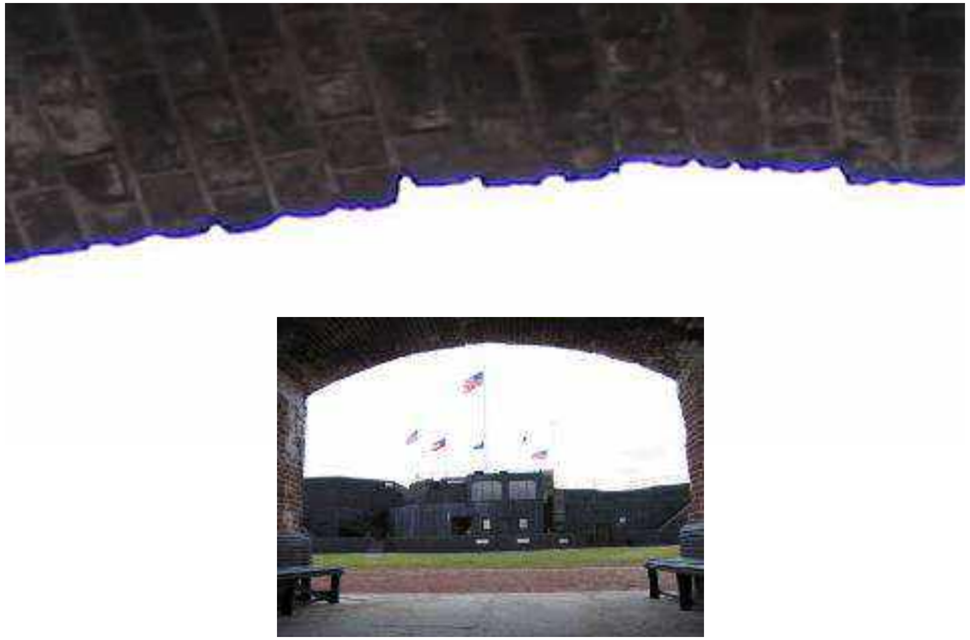
Geometric Distortions



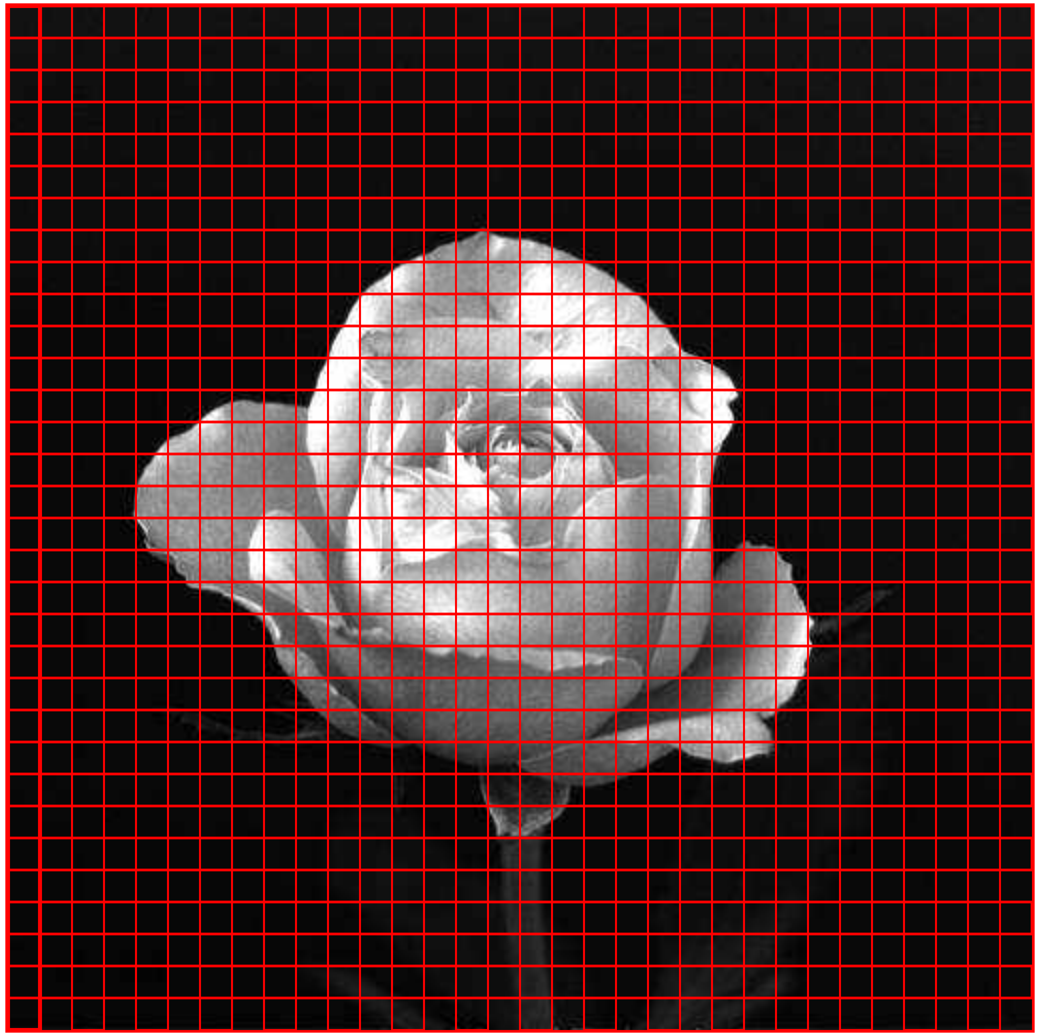
Chromatic Aberration

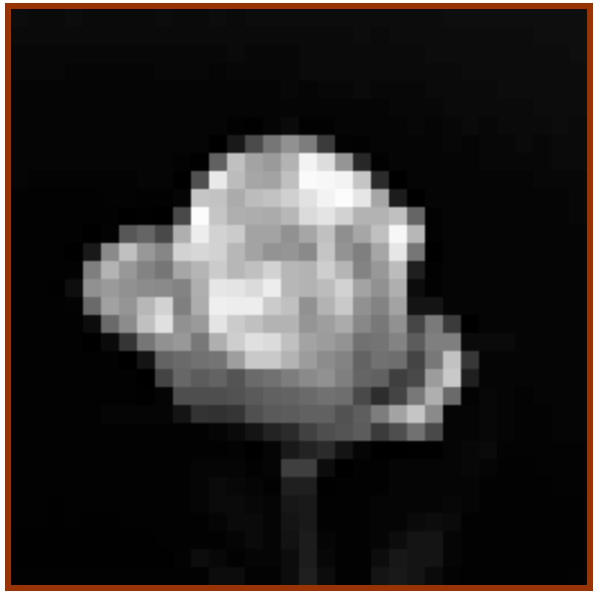
Normal lenses diffract different wavelengths to different degree



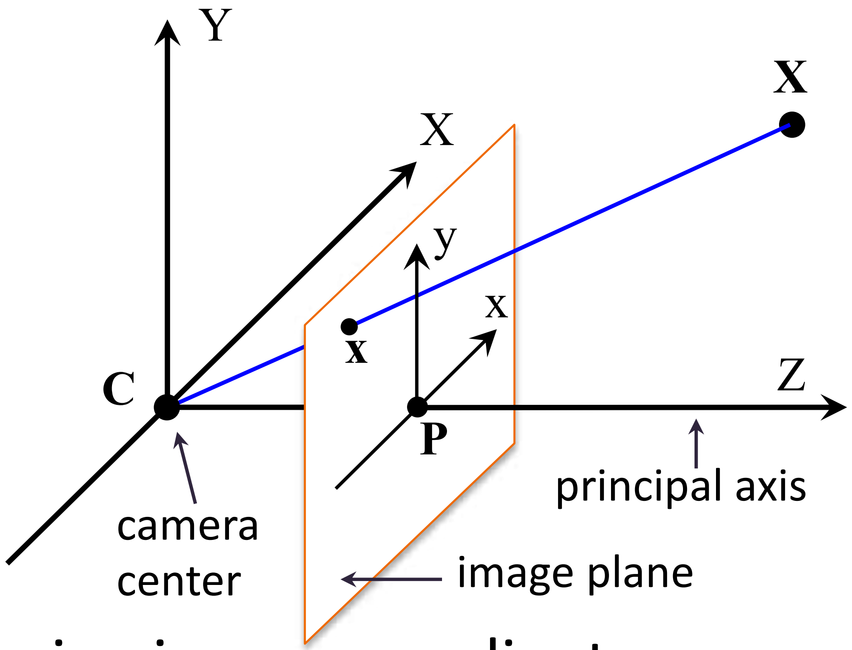


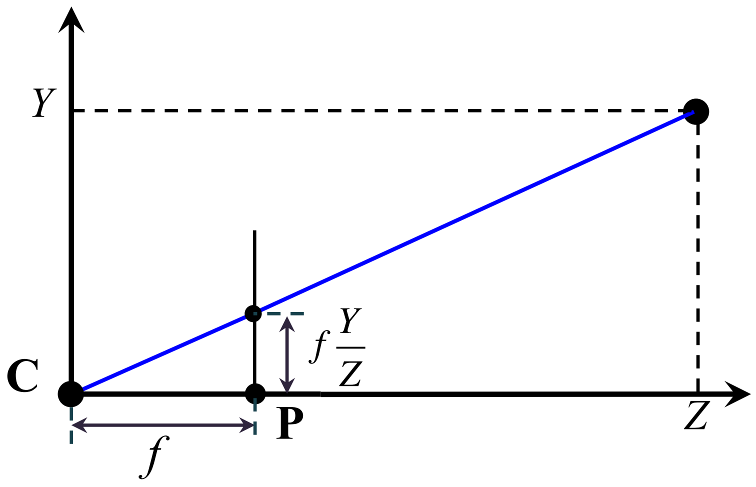
Sampling an Image: Resolution





The Camera: A Mathematical Model





- Cartesian image coordinates:

$$x = f\frac{X}{Z}, \qquad y = f\frac{Y}{Z}$$

- In matrix form (homogeneous):

$$\mathbf{x} = \begin{bmatrix} x \\ y \\ w \end{bmatrix} = \begin{bmatrix} f & 0 & 0 & 0 \\ 0 & f & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} X \\ Y \\ Z \\ 1 \end{bmatrix} = \mathbf{PX}$$

Note:
- Camera at origin, Z axis along look vector
- Orthogonal Image axes
- Uniform scale

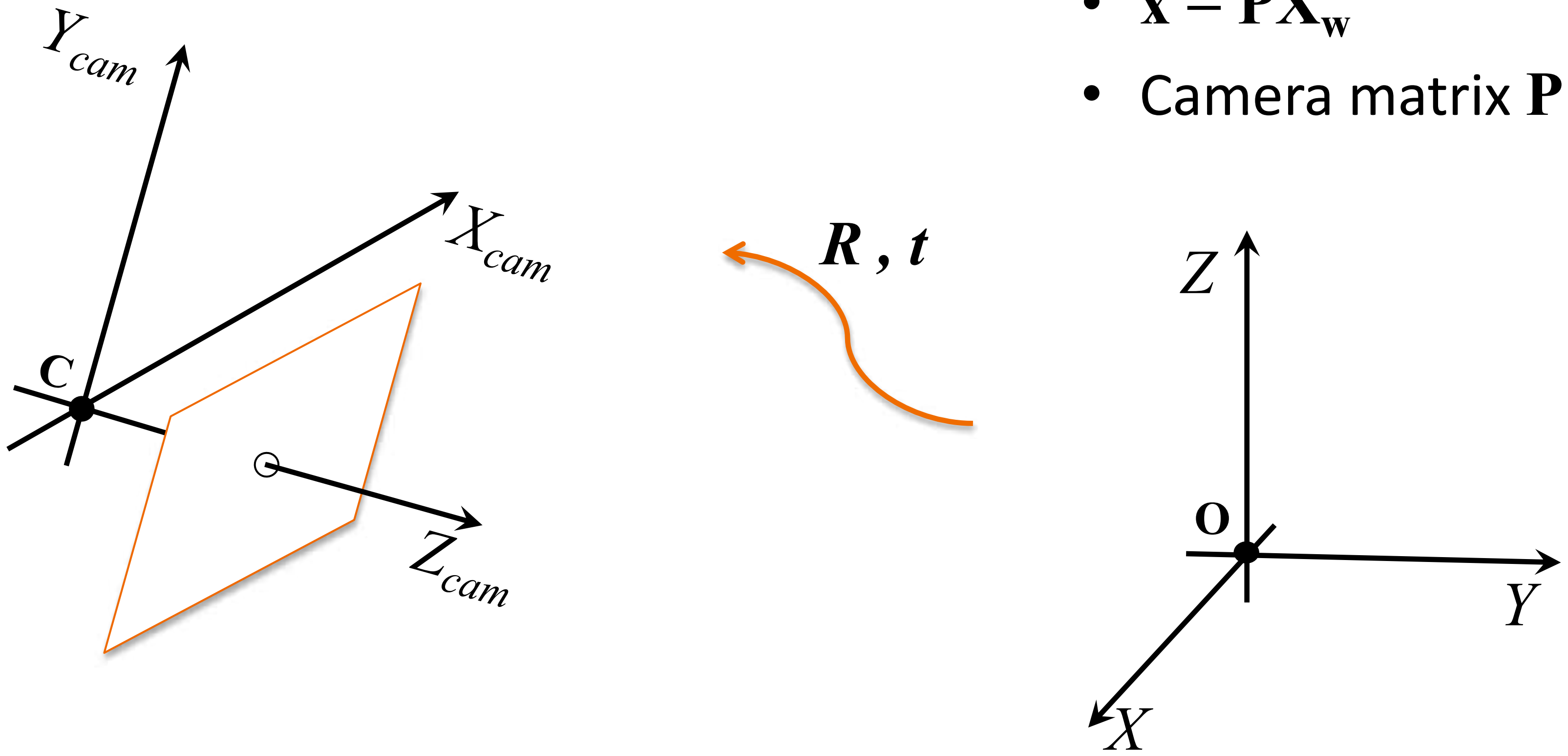
Moving the Camera from Origin

- General Setting: Camera is not at origin and Z is not the optical axis.
- Camera is at a point C in world coordinates. The camera axes are also rotated by a matrix R.

In General,

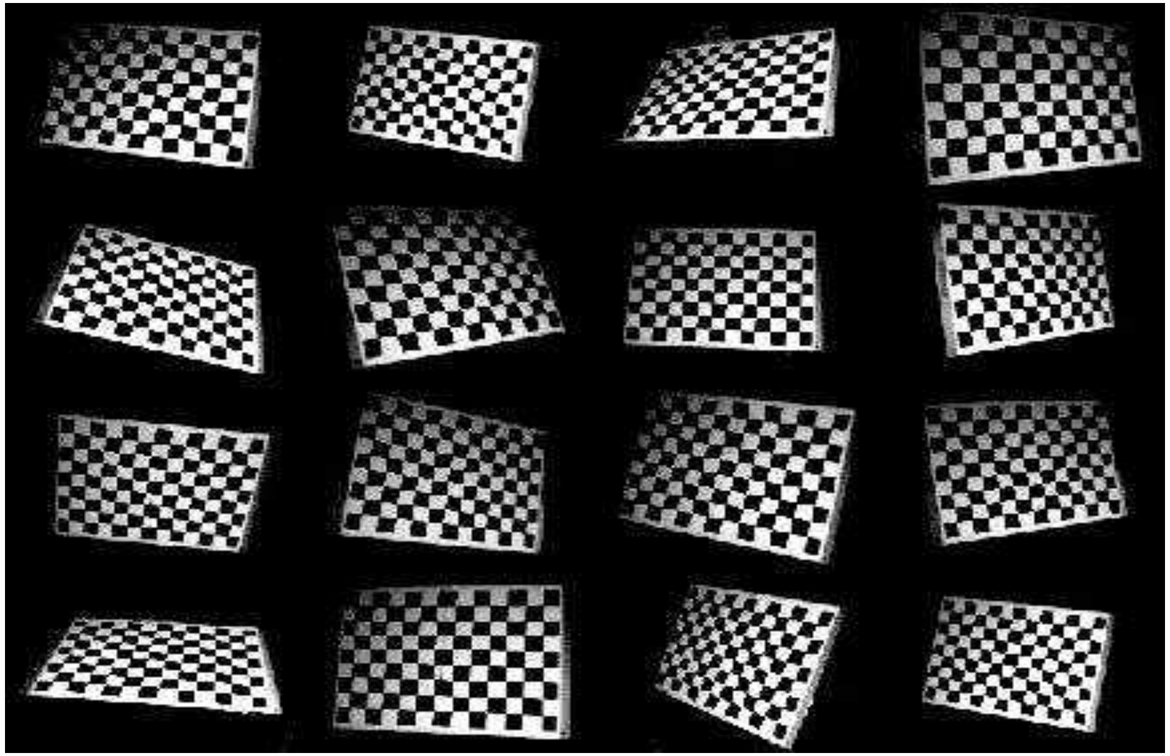
- $\mathbf{x} = \mathbf{P}\mathbf{X}_w$
- Camera matrix $\mathbf{P} = \mathbf{K}\,[\mathbf{R} \mid \mathbf{t}]$

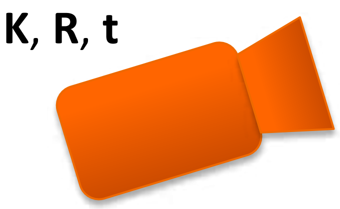
Calibration



Calibration Methods

1. 3D Reference Object based calibration
2. Calibration from a precisely moving plane (R.Y. Tsai)
3. Calibration using a plane with unknown motion

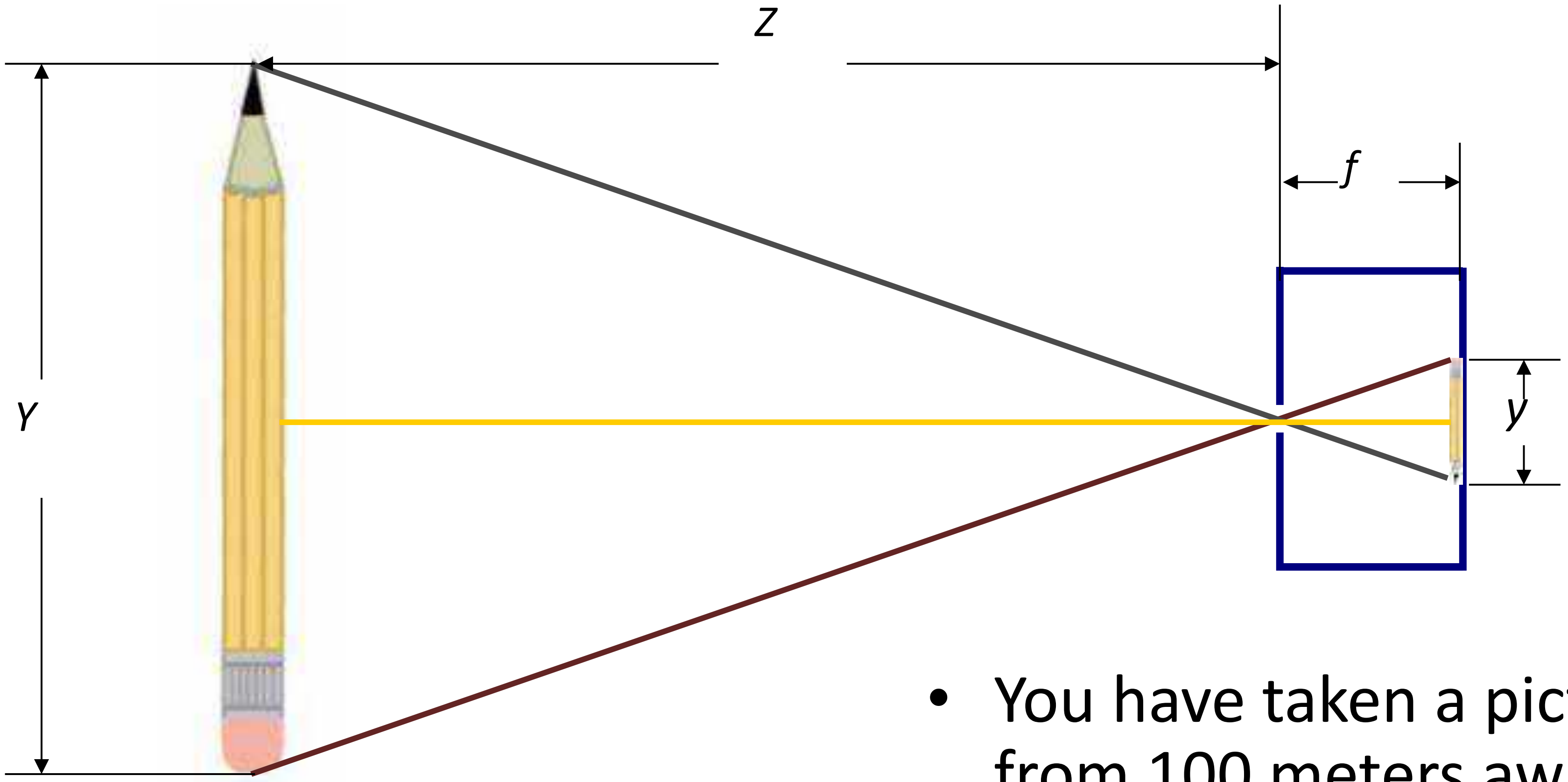




Outline

- Introduction to Computer Vision [20 minutes]
  – What, why, why not?
- Camera Model and Geometry [20 minutes]
- Problems in Computer Vision
  – Recovering world geometry [20 minutes]
    - 4 Geometric Properties and their applications
  – Reorganizing images [20 minutes]
  – Detection and Recognition [20 minutes]
- Questions and Discussions [10 minutes]

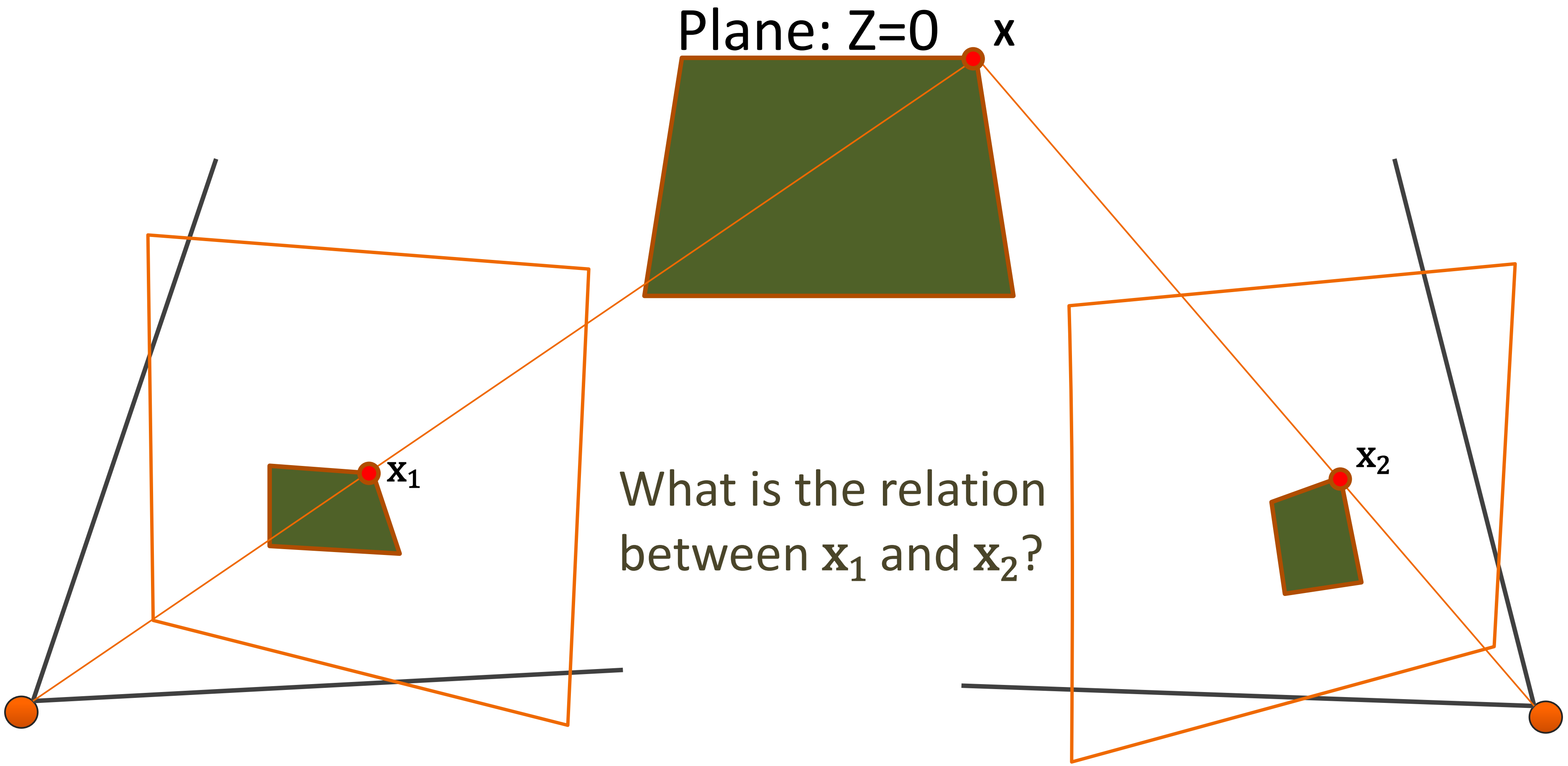
The Pinhole Camera



$$y = f\frac{Y}{Z}$$

- You have taken a picture of a building from 100 meters away using a camera of focal length 20mm. The height of the image is 10mm.
- What is the height of the building?

Two-View Geometry: Planar World



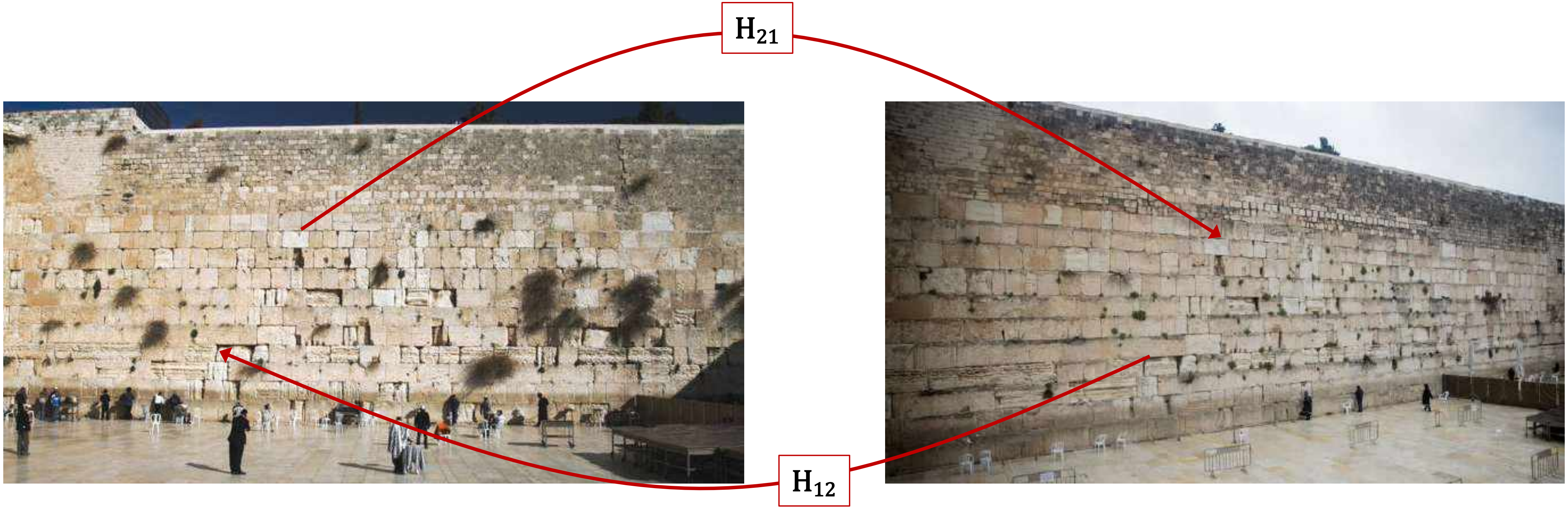
What is the relation between $\mathbf{x}_1$ and $\mathbf{x}_2$?

$$\boxed{\mathbf{x}_1 = \mathbf{H}_{12}\mathbf{x}_2; \qquad \mathbf{x}_2 = \mathbf{H}_{21}\mathbf{x}_1}$$

where H is a 3 × 3 non-singular matrix

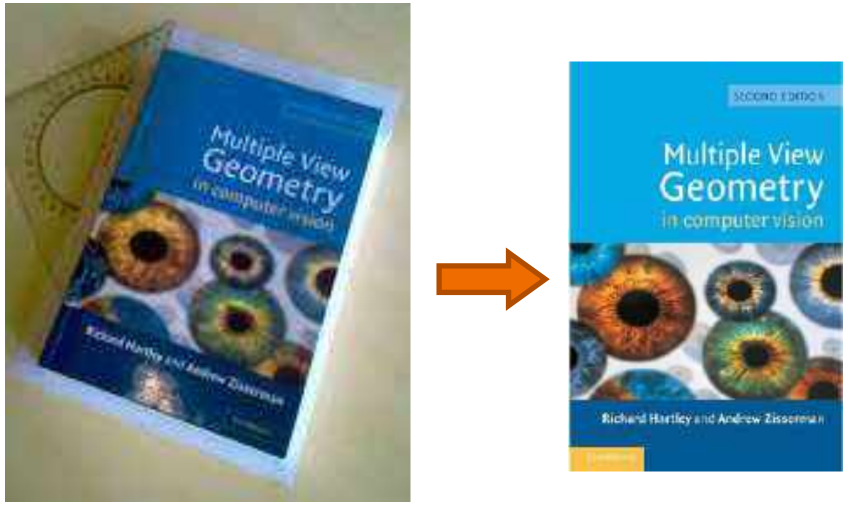
Planar Homography

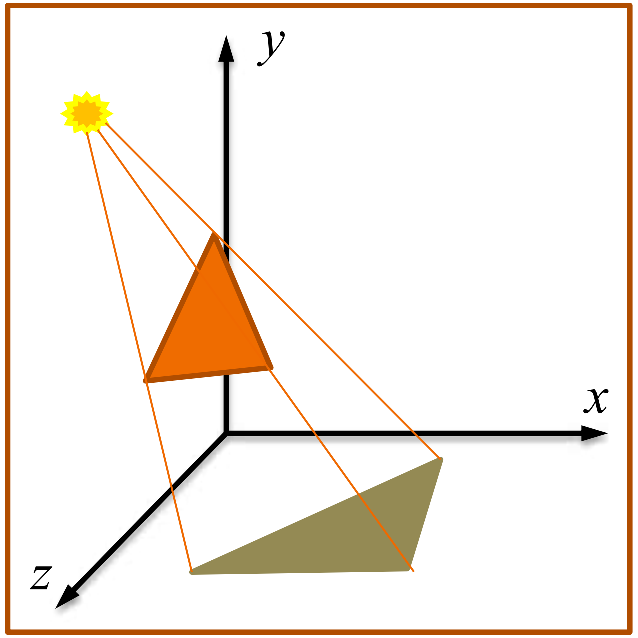
- Given two images of a planar world, every pixel in an image can be computer from the other. Its location is given by $\mathbf{x}_a = \mathbf{H}\mathbf{x}_b$

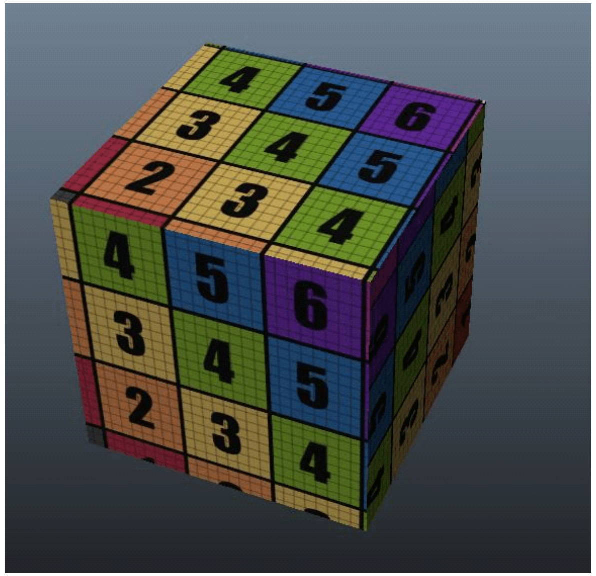


Planar Homography: Applications

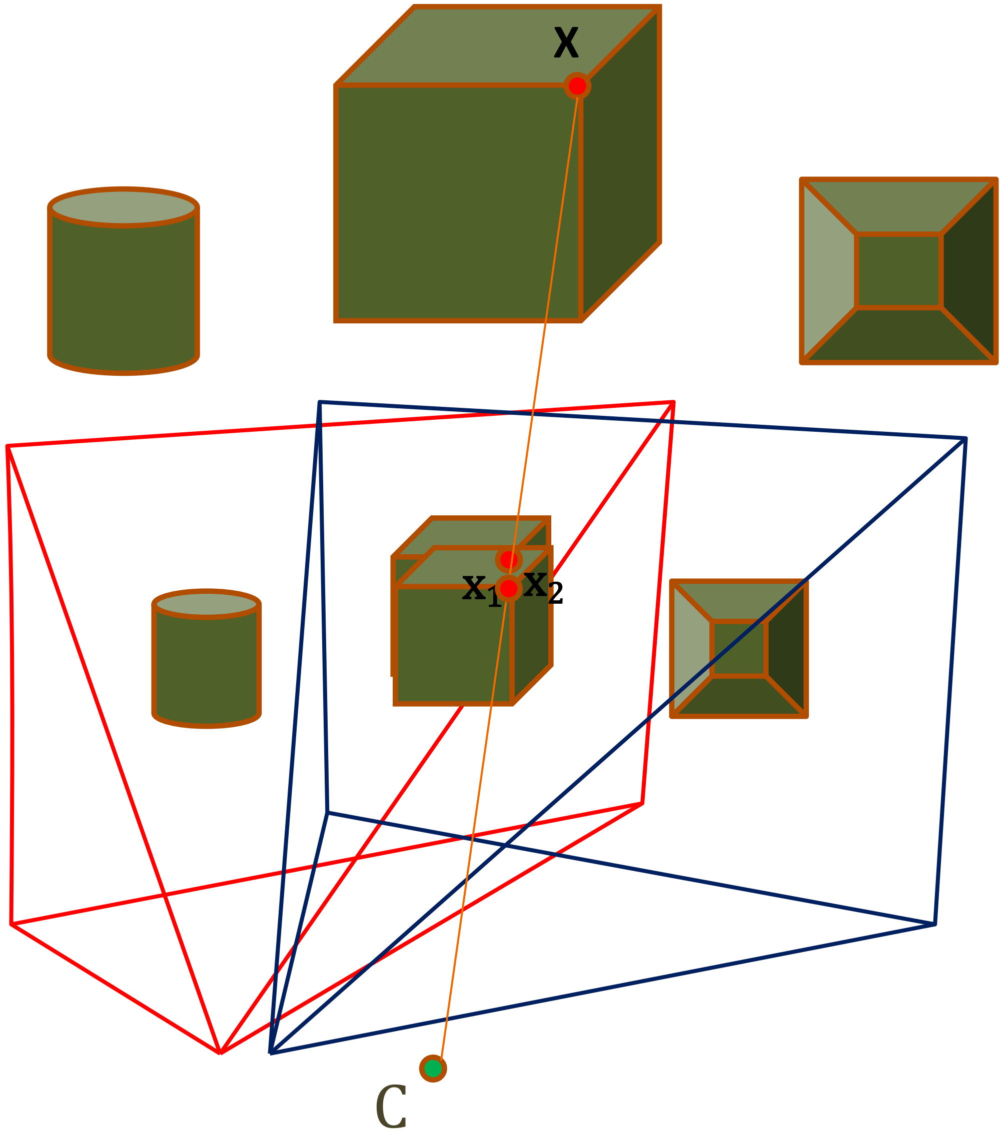
- Removing perspective distortion
- Rendering planar textures
- Rendering planar shadows
- Estimating Camera Pose; AR







Two-View; Case 2: Same Camera Center



Arbitrary world

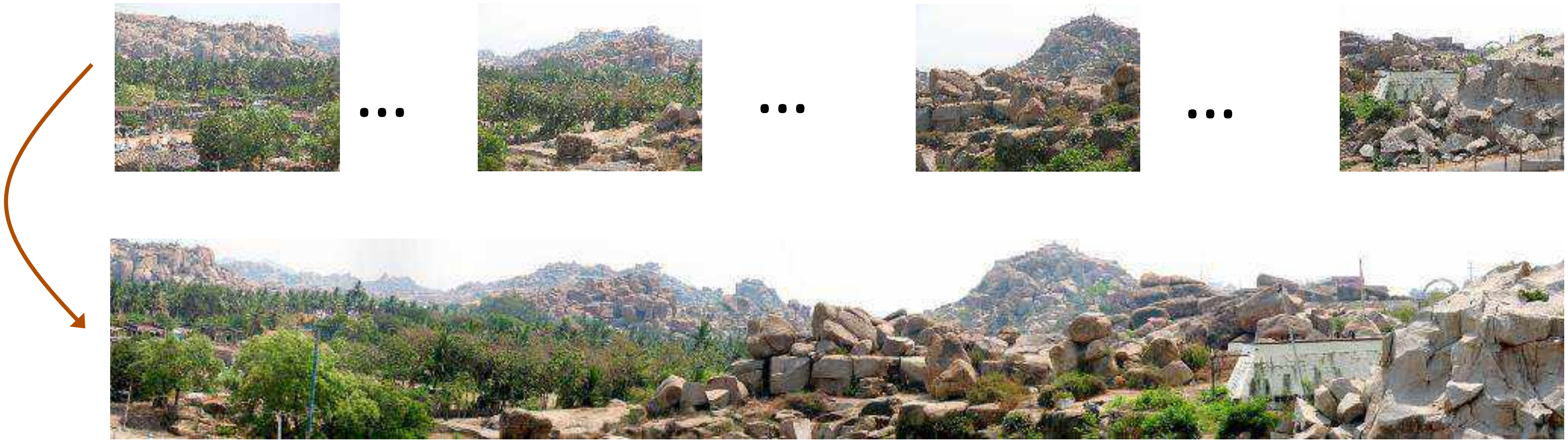
What is the relation between $\mathbf{x}_1$ and $\mathbf{x}_2$?

$$\boxed{\mathbf{x}_1 = \mathbf{H}_{12}\mathbf{x}_2; \qquad \mathbf{x}_2 = \mathbf{H}_{21}\mathbf{x}_1}$$

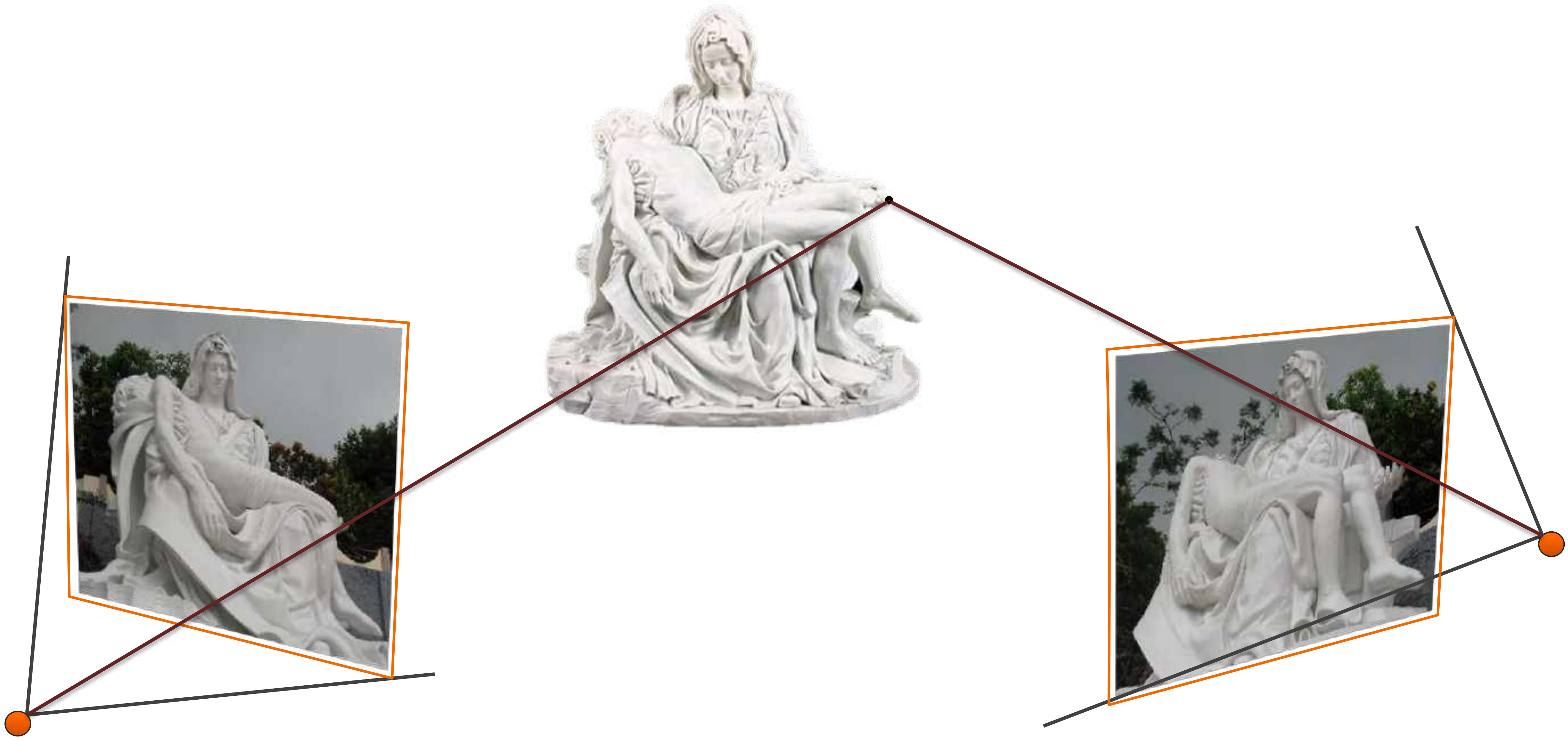
where H is a 3×3 non-singular matrix

Homography: Applications

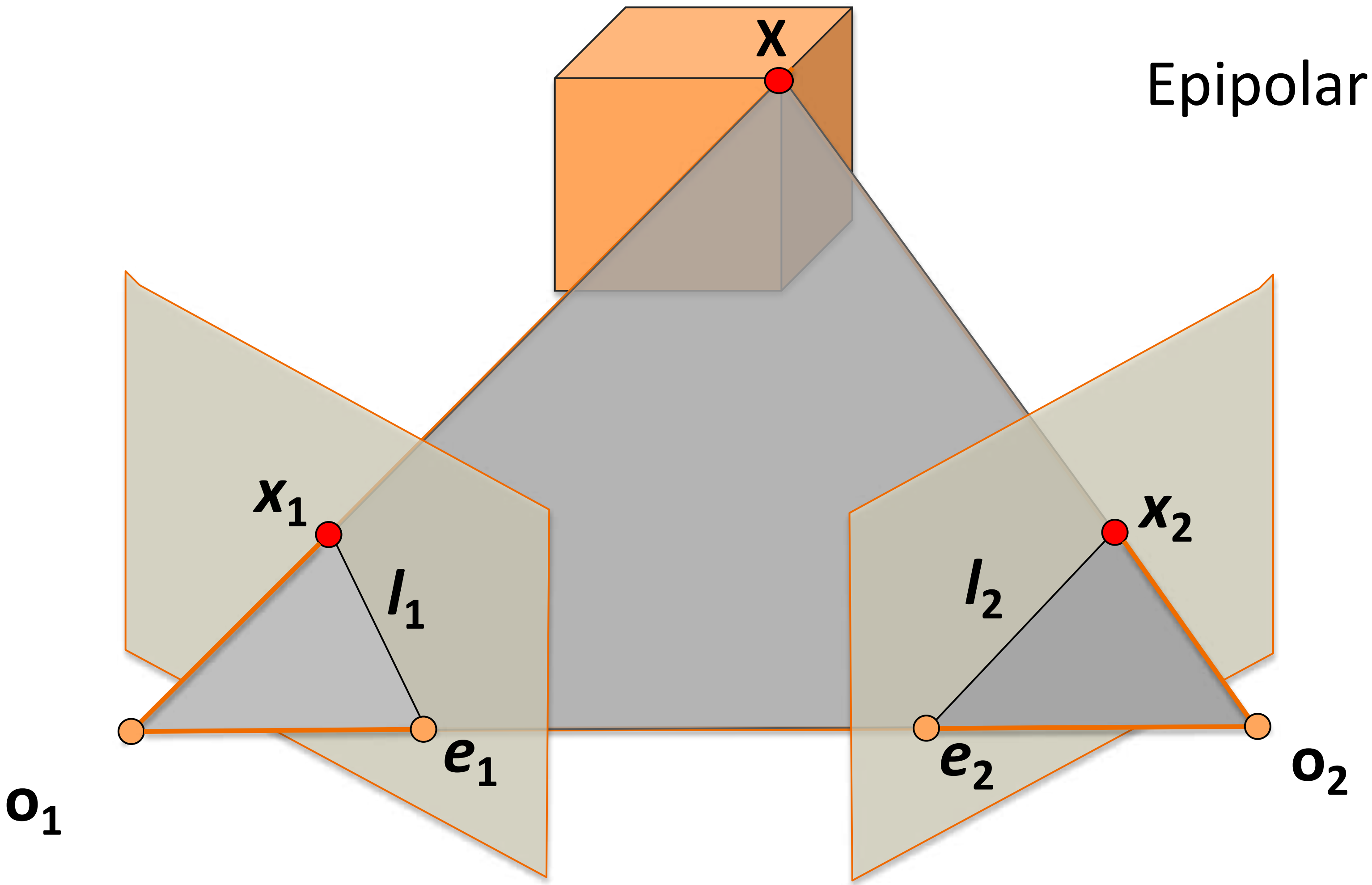
- Image Mosaicing
- Detecting camera translation
- Multi-frame Super-resolution



Case 3: Generic 3D World and Cameras

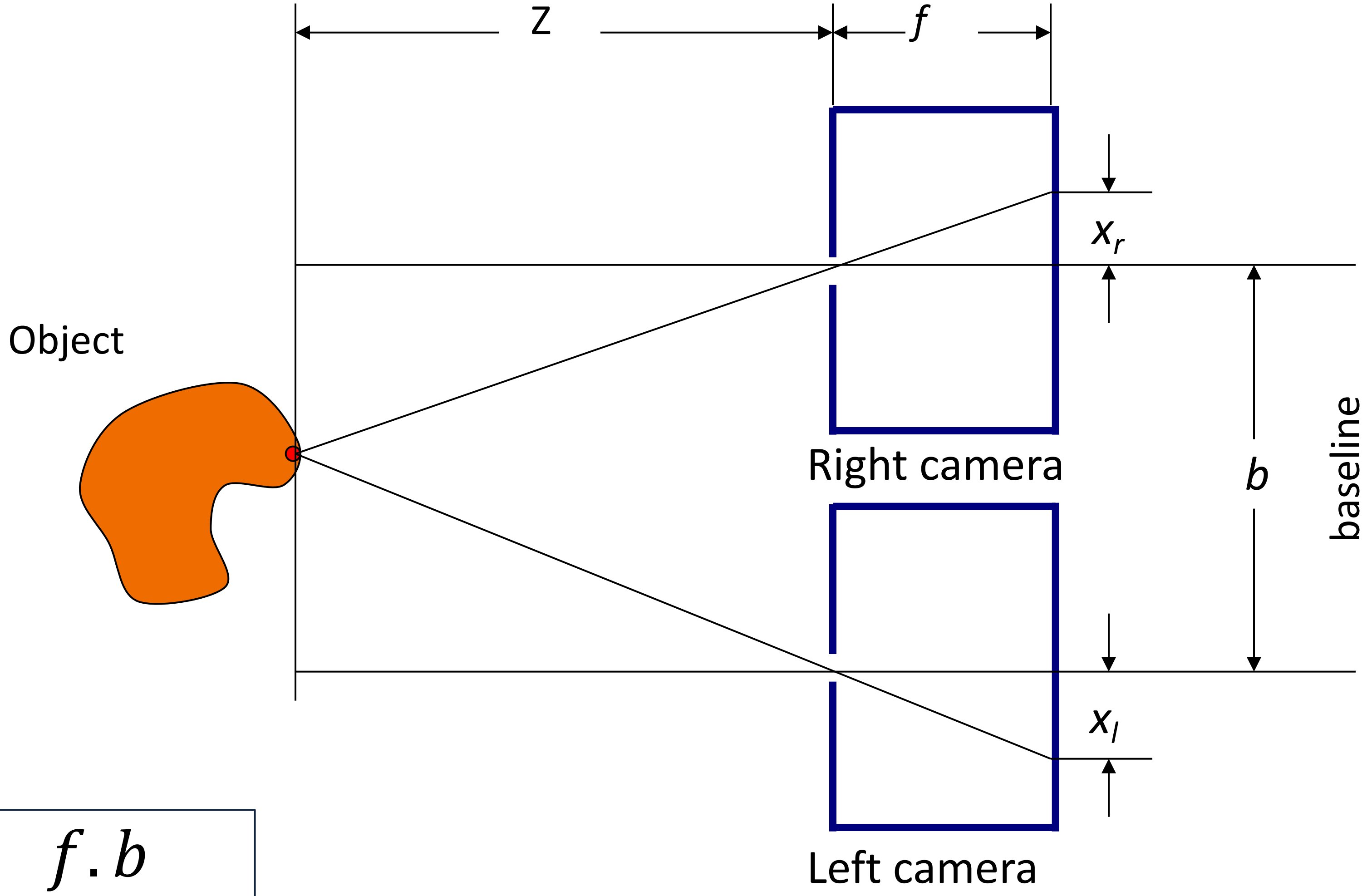


Case 3: Generic 3D World and Cameras



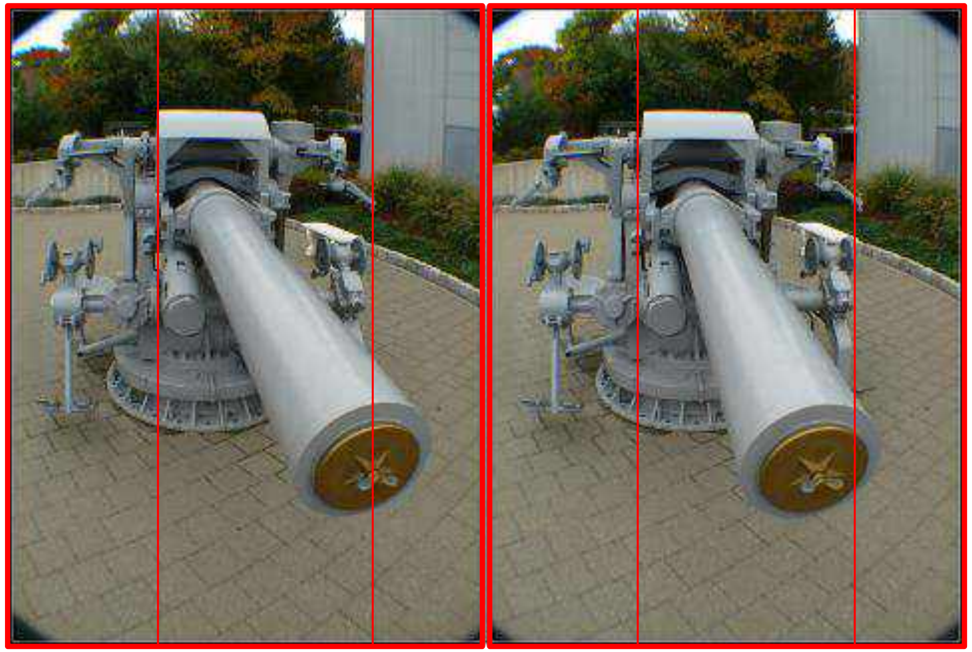
Epipolar Geometry

Stereo



$$Z = \frac{f.b}{(x_r + x_l)}$$

Stereo Geometry



- Farther the point, smaller the disparity and vice versa
- A large baseline can give more reliable estimates of depth. However, matching becomes harder
- Basic step: Identify common points in the two camera views
  – How do we find the images of a single world point in two views?
  – Search for similar appearance
  – Use Epipolar Geometry to reduce search

Outline

- Introduction to Computer Vision [20 minutes]
  - What, why, why not?
- Camera Model and Geometry [20 minutes]
- Problems in Computer Vision
  - Recovering world geometry [20 minutes]
  - Reorganizing images [20 minutes]
  - Detection and Recognition [20 minutes]
- Questions and Discussions [10 minutes]

Thresholding

Decide each pixel to be part of an object or background depending on its gray value

$$t(m,n) = \begin{cases} 1 & \text{if } u(m,n) > T \\ 0 & \text{if } u(m,n) \leq T \end{cases}$$

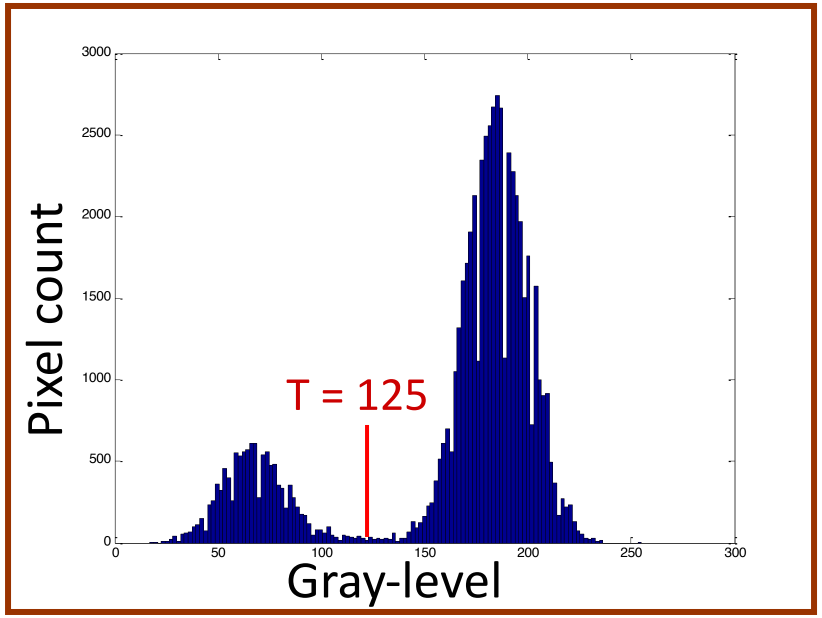
Original

Thresholded (T=95)

Histogram and Segmentation

- A count of pixels of each graylevel (or range of graylevels) in an image

Grayscale Image



Histogram

Thresholded (T=125)

Automatic Thresholding

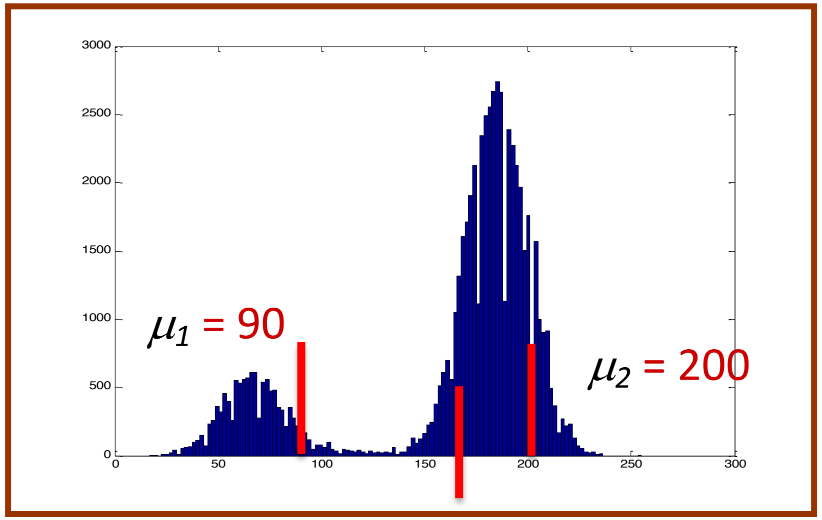
1. Select an initial estimate of T

2. Segment the image using T. Compute the mean gray values of the regions, μ_1 and μ_2

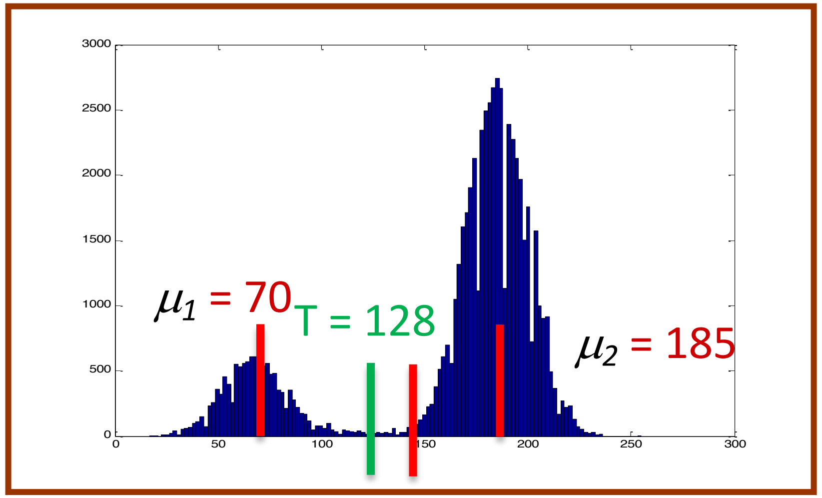
3. Set the new threshold $T=(\mu_1+\mu_2)/2$

4. Repeat 2 and 3 until T stabilizes

Assumptions: normal distribution, low noise

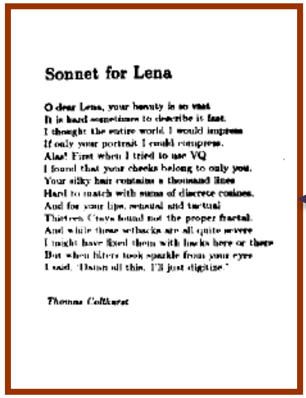


T = 170

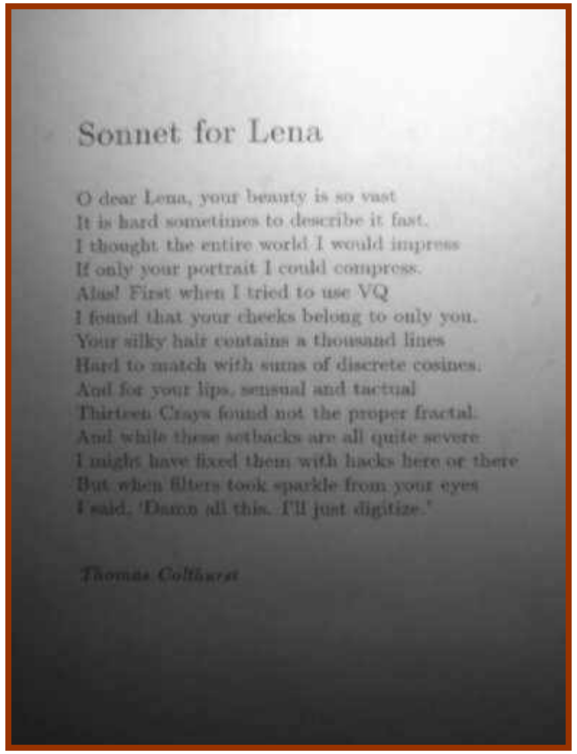


T = 145

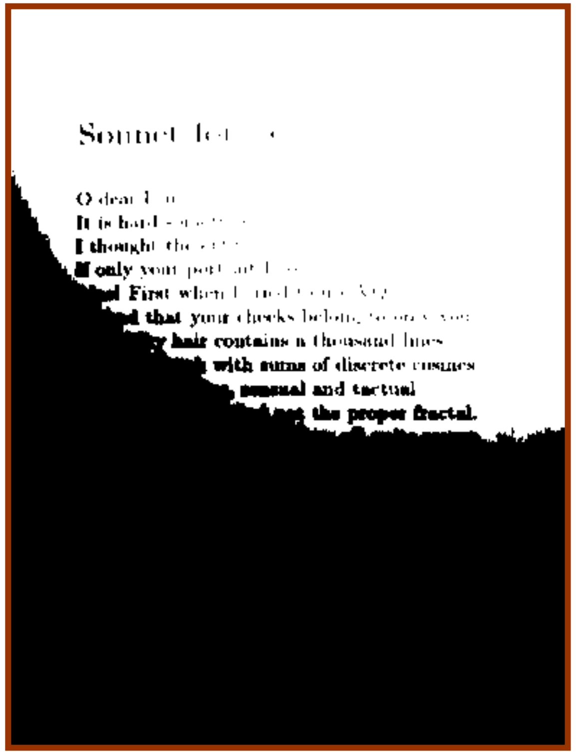
Adaptive Thresholding



e.g., Chow & Kaneko Thresholding:

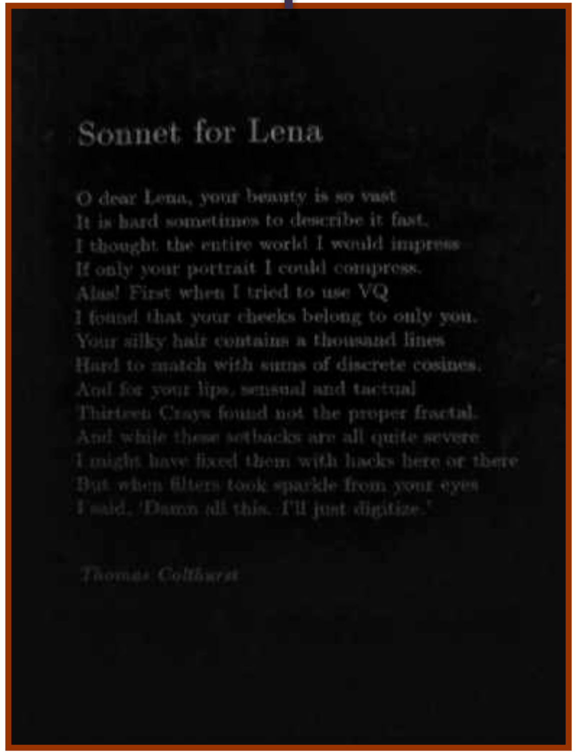


Original



Single Threshold

Low-pass filtered

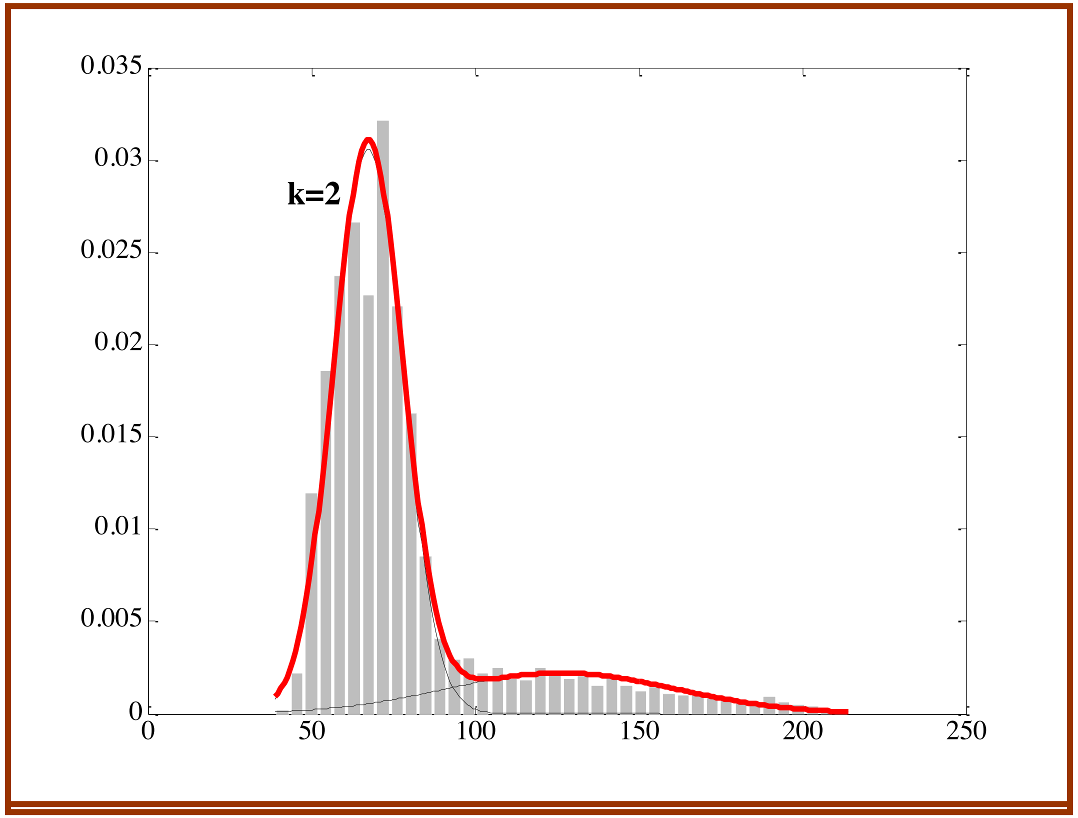


Difference

Optimal Thresholding

- The graylevel histogram is approximated using a mixture of two gaussians and threshold chosen to minimize the segmentation error

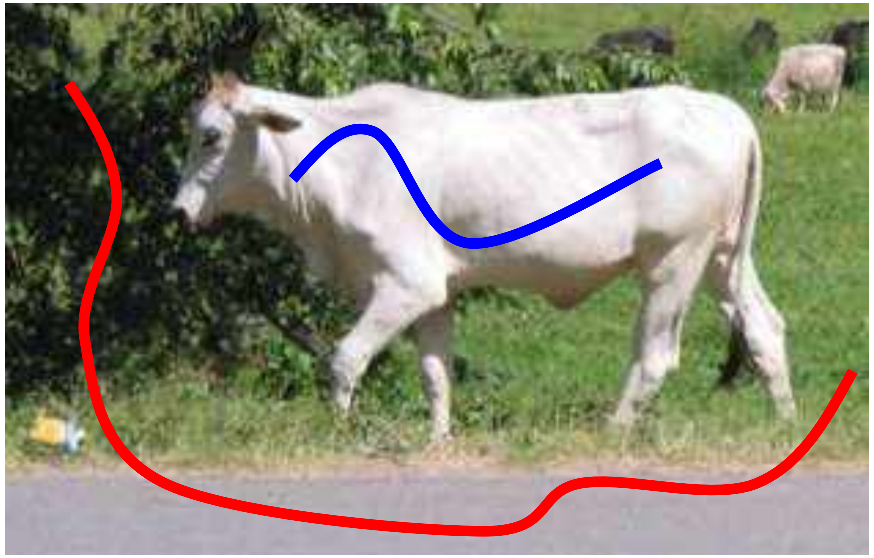
Grayscale Image



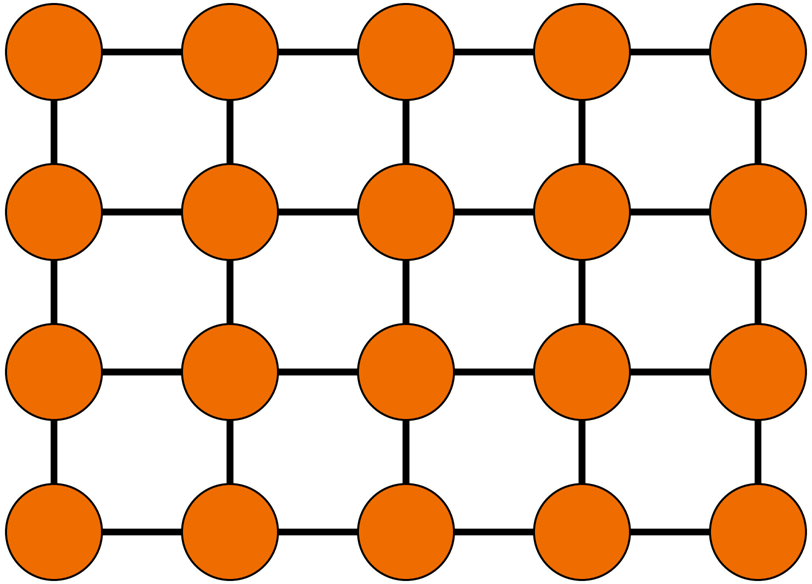
Histogram with bimodal fit

Thresholded (T=94)

Graph Cuts for Binary Image Segmentation



Object - white, Background - green/grey



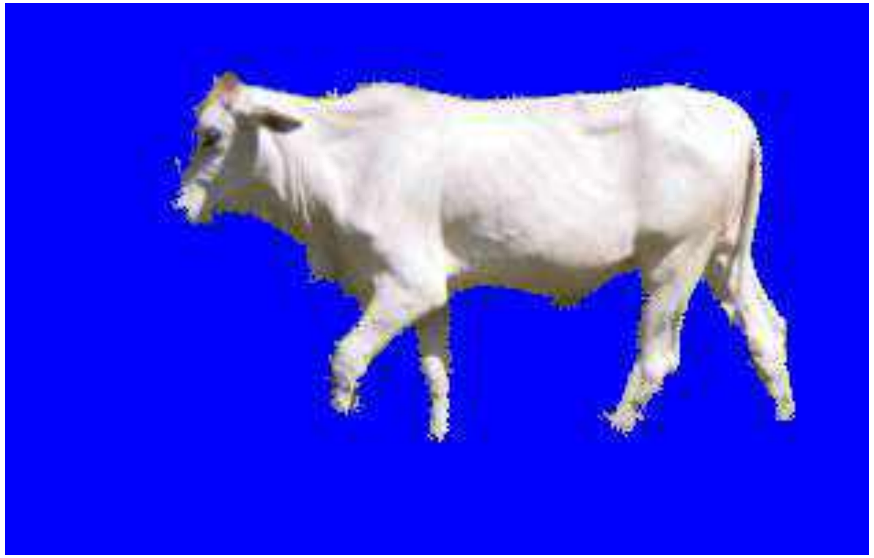
Graph G = (V,E)

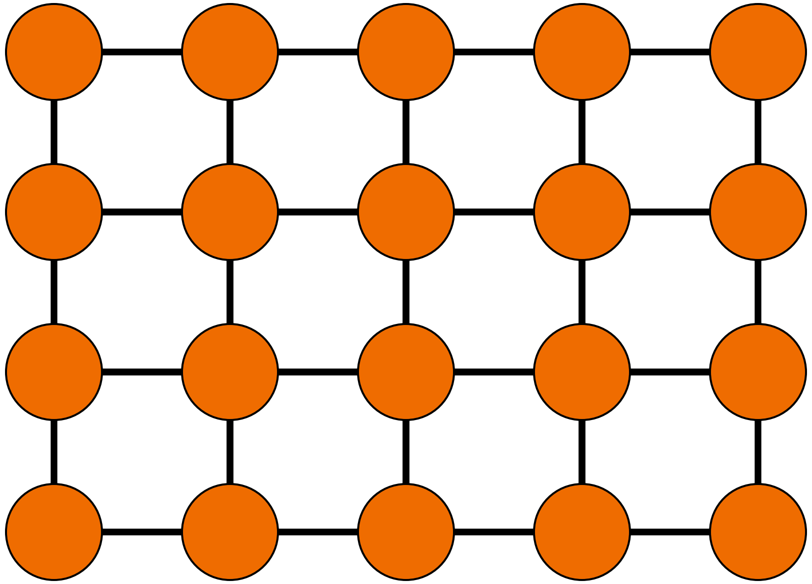
$$Q(f;\theta) = \sum_a \theta_{a;f(a)} + \sum_{(a,b)} \theta_{ab;f(a)f(b)}$$

Pairwise Potential

Problem: Find the labeling with minimum cost f*

Graph Cuts for Binary Image Segmentation





Graph G = (V,E)

L = {fg, bg}

Vertex corresponds to a pixel

Edges define grid graph

Several other methods to optimize over the graph (MRF)

GrabCuts: An Intelligent Extension

Fast &
Accurate ?

- Less user input: only rectangle
- Handle color
- Extract matte as post-process

“Grabcut: Interactive Foreground Extraction using Iterated Graph Cuts”, by Rother, Kolmogorov and Blake, Siggraph 2004.

Approach: Iterated Graph Cuts

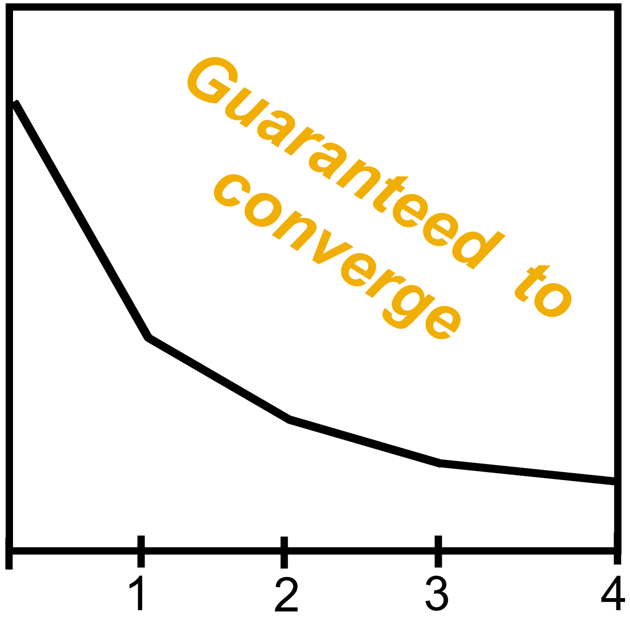
User Initialisation

Learn foreground color model

Graph cuts to infer the foreground

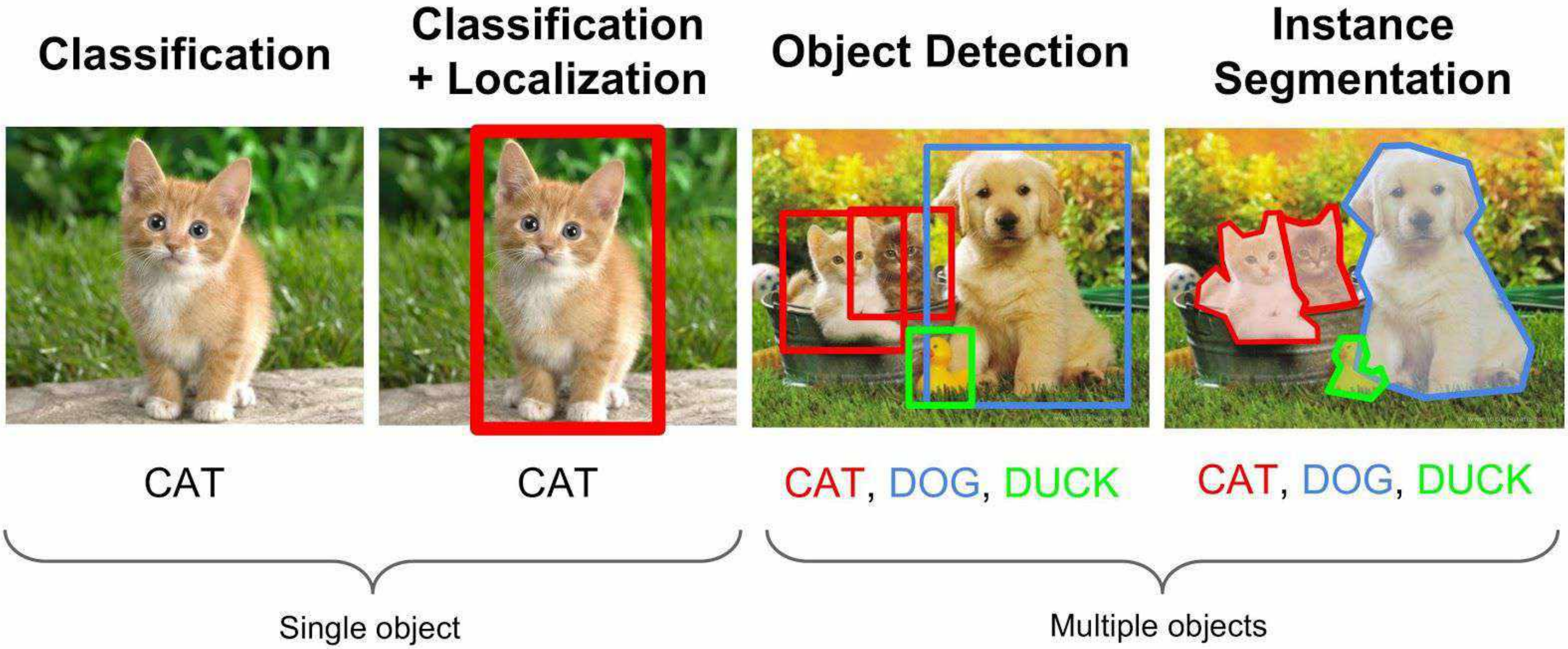
Iterated Graph Cuts

Result



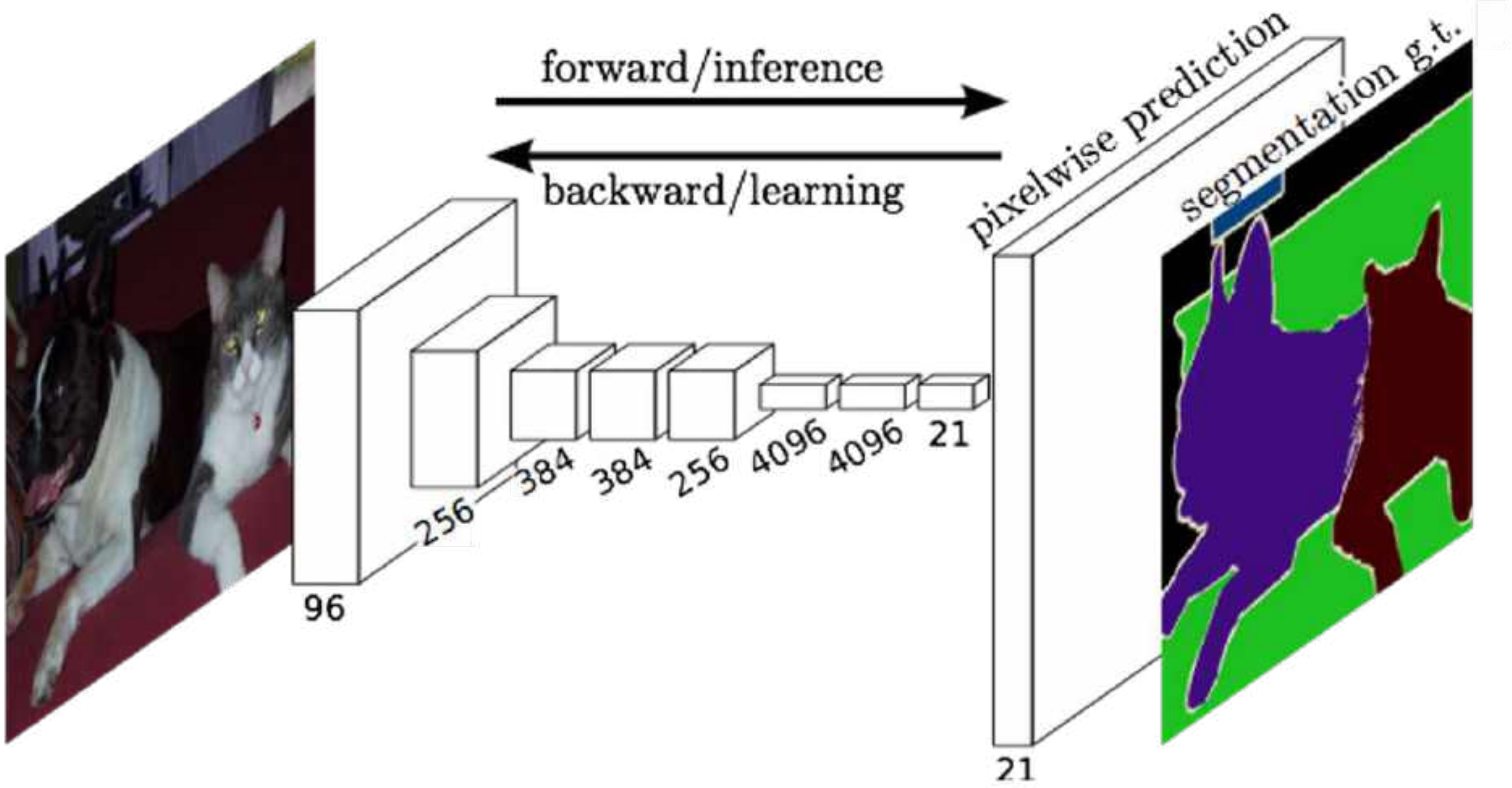
Energy after each Iteration

Semantic and Instance Segmentation



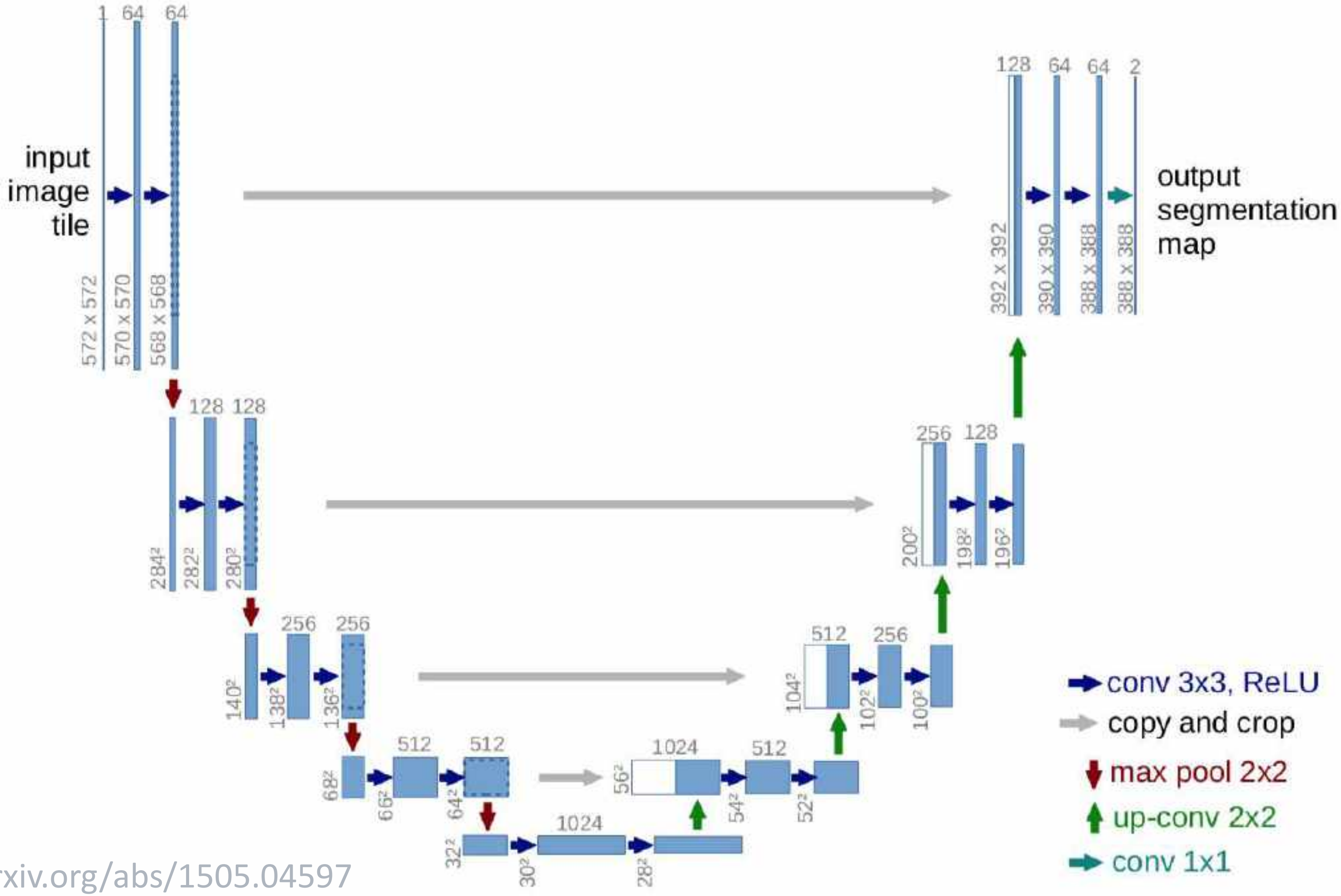
Source http://cs224d.stanford.edu/index.html

FCN for Semantic Segmentation



- Challenge: Segmentation = Classification + Localization
- Classification needs larger context and location invariance
- Localization needs sensitivity to location

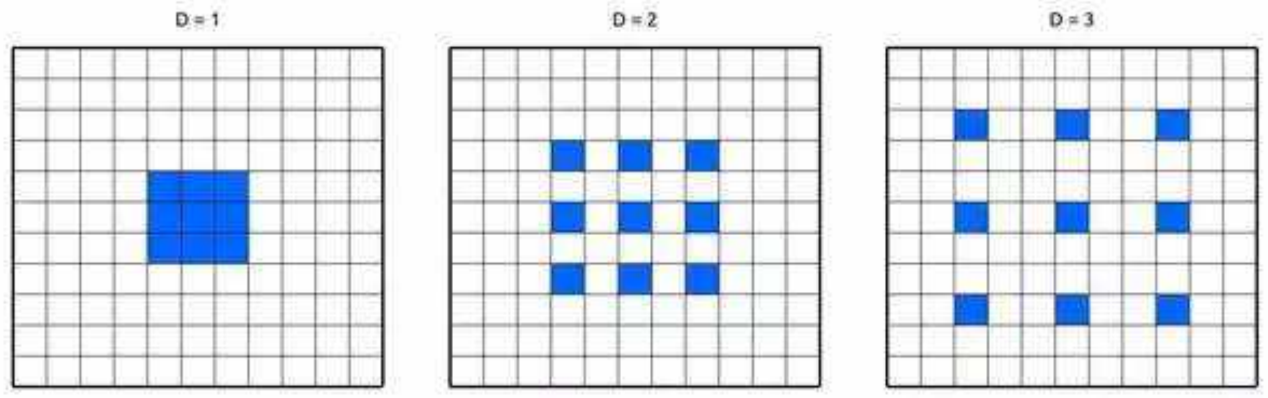
Unet: Skip Connections



Source:- https://arxiv.org/abs/1505.04597

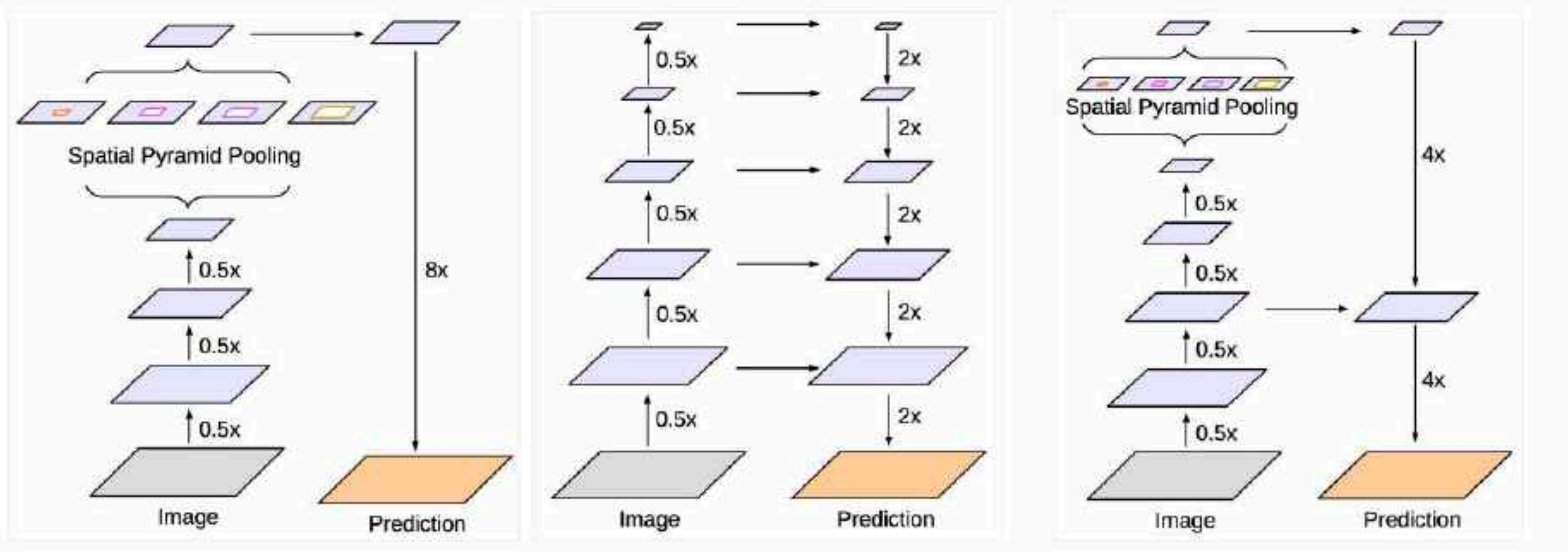
DeepLab V3+: Dilation, Spatial Pyramids, Depthwise Sep.

Atrous Convolutions



Spatial Pyramids + Enc-Dec + Atrous Conv.

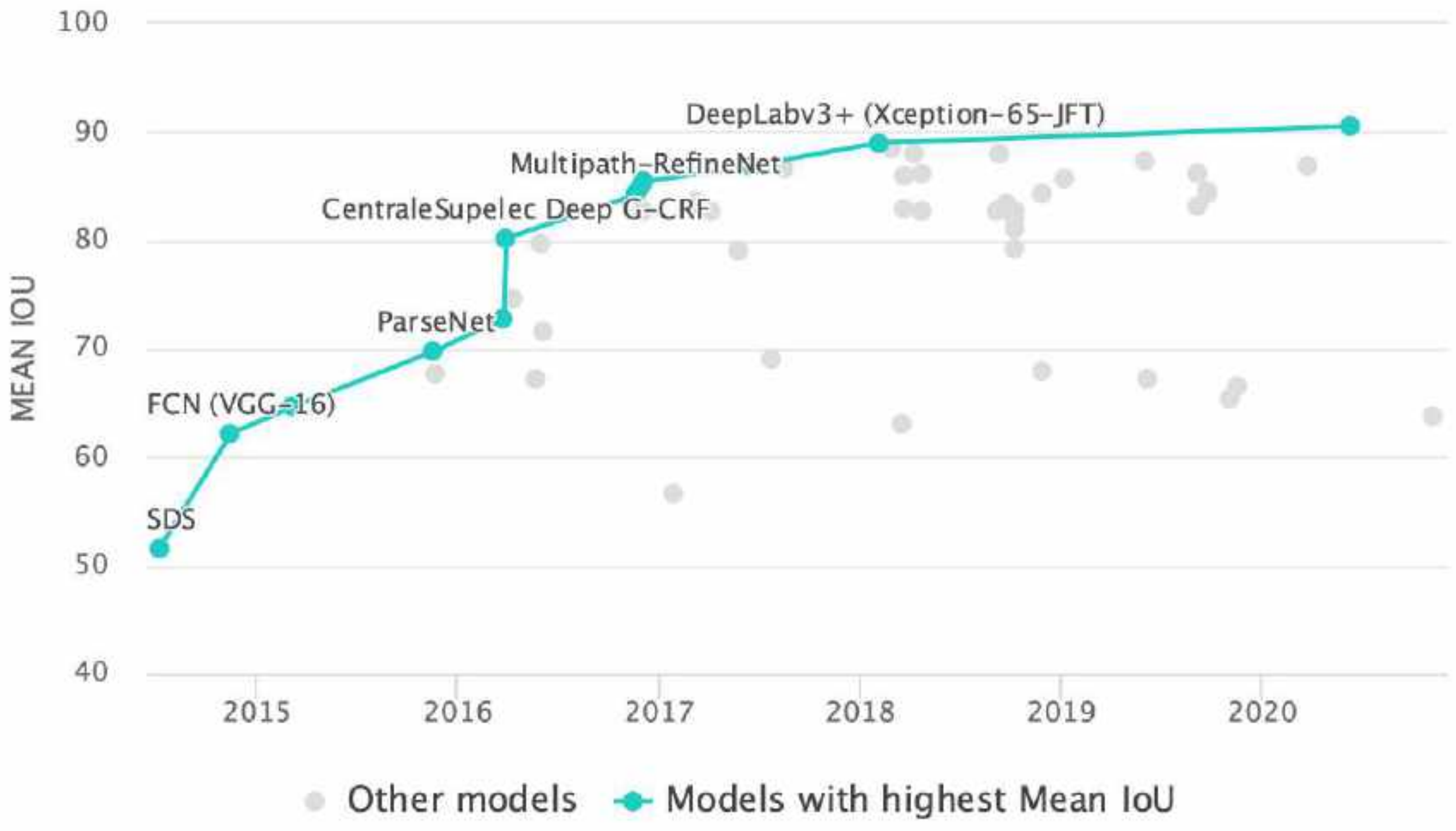
Depth-wise Separable Conv. (Xception Model)



(a) Spatial Pyramid Pooling (b) Encoder-Decoder (c) Encoder-Decoder with Atrous Conv

Evolution of Segmentation

- GCN: Larger Kernels, Boundary Refinement
- CRF for regularization
- Better Training
  - Self Training
- Use of LSTM for video segmentation



Source: https://paperswithcode.com/sota/semantic-segmentation-on-pascal-voc-2012

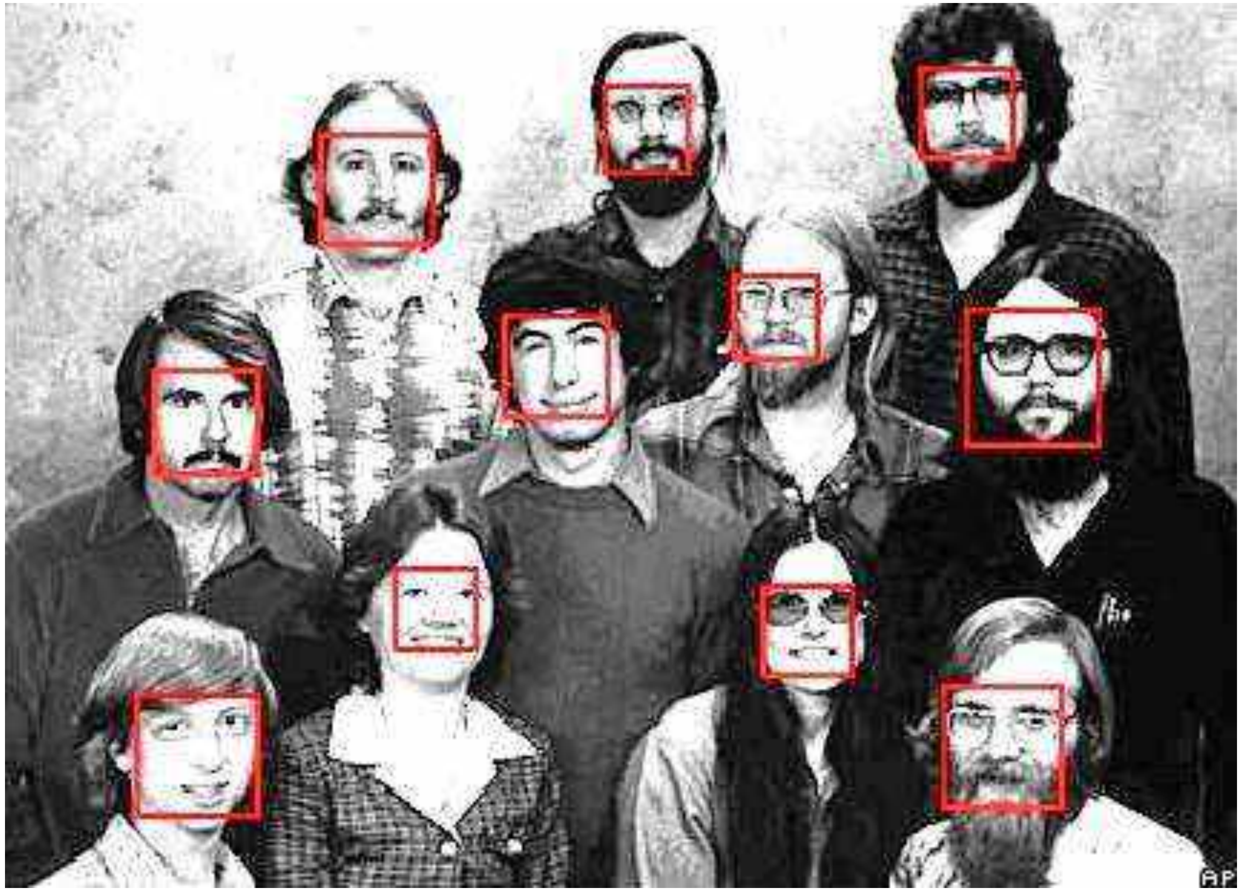
Applications

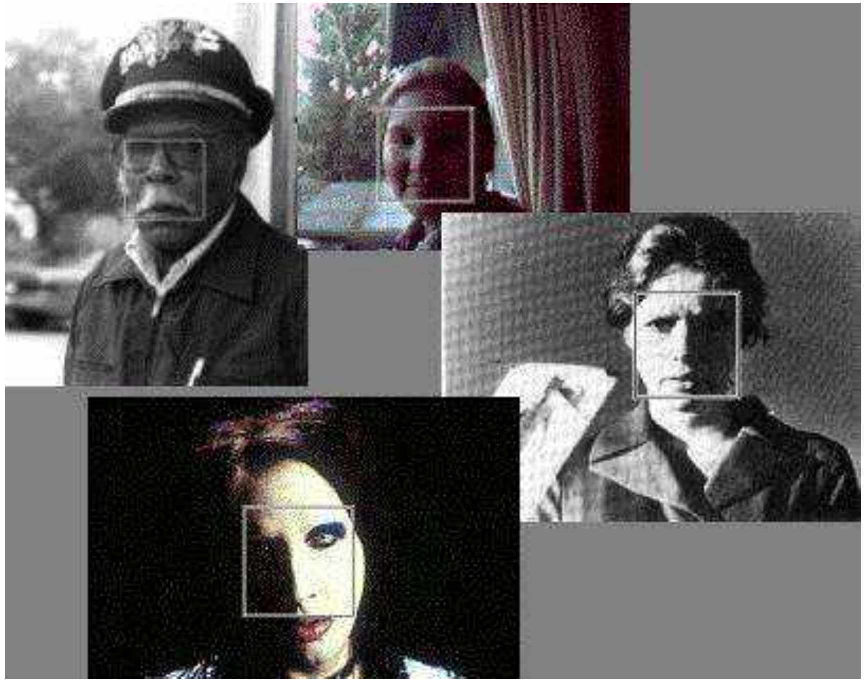
- Portrait Mode
- Background removal in online meetings
- Virtual Makeup, Virtual Try-on
- Monocular Depth Estimation
  - Autonomous Navigation
  - Map Generation
- Image Editing

Outline

- Introduction to Computer Vision [20 minutes]
  - What, why, why not?
- Camera Model and Geometry [20 minutes]
- Problems in Computer Vision
  - Recovering world geometry [20 minutes]
  - Reorganizing images [20 minutes]
  - Detection and Recognition [10 minutes]
- Questions and Discussions [15 minutes]

The Task: Say Face Detection





Approach 1: Classify Each Window

Slide a window across image and evaluate a face model at every location

How Many Windows (Speed)

- 1280×1024 image; 24×24 to 1024×1024 windows

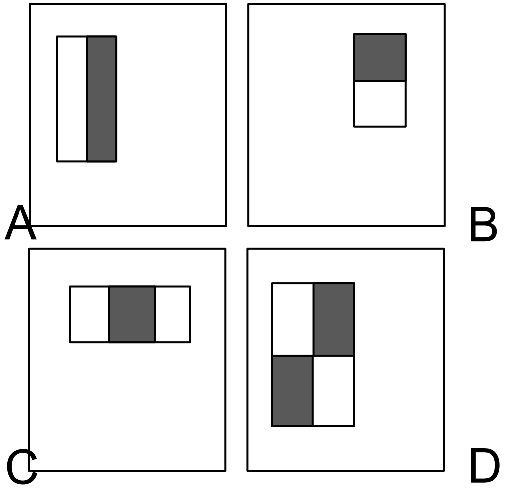
- # of Windows?
  - ~ 1 million locations
  - 18 scales/window sizes at 1.25 multiples
  - 14.5 million potential face candidates
  - Features to be extracted for each candidate window

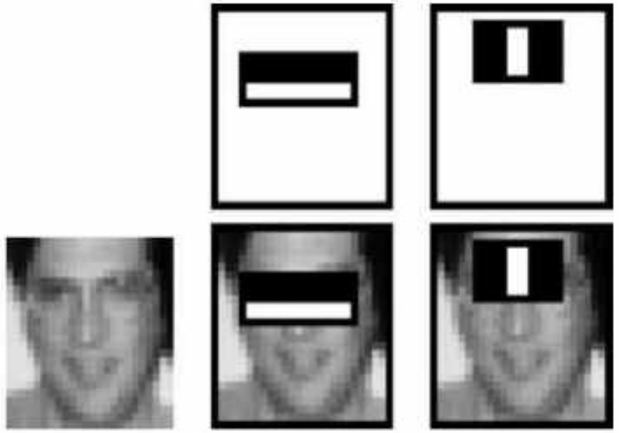
Accuracy (FPR)

- Classify each as face or non-face
  - What should be the accuracy (False Positive Rate)?
- Faces are rare: 0-10 per image
  - For computational efficiency, we should try to spend as little time as possible on the non-face windows
  - To avoid having a false positive in every image, the false positive rate has to be less than $\mathbf{10^{-6}}$

The Viola/Jones Face Detector

- A seminal approach to real-time object detection
- Key ideas
  - *Haar Features + Integral images* for fast feature evaluation
  - *Attentional cascade* for fast rejection of non-face windows

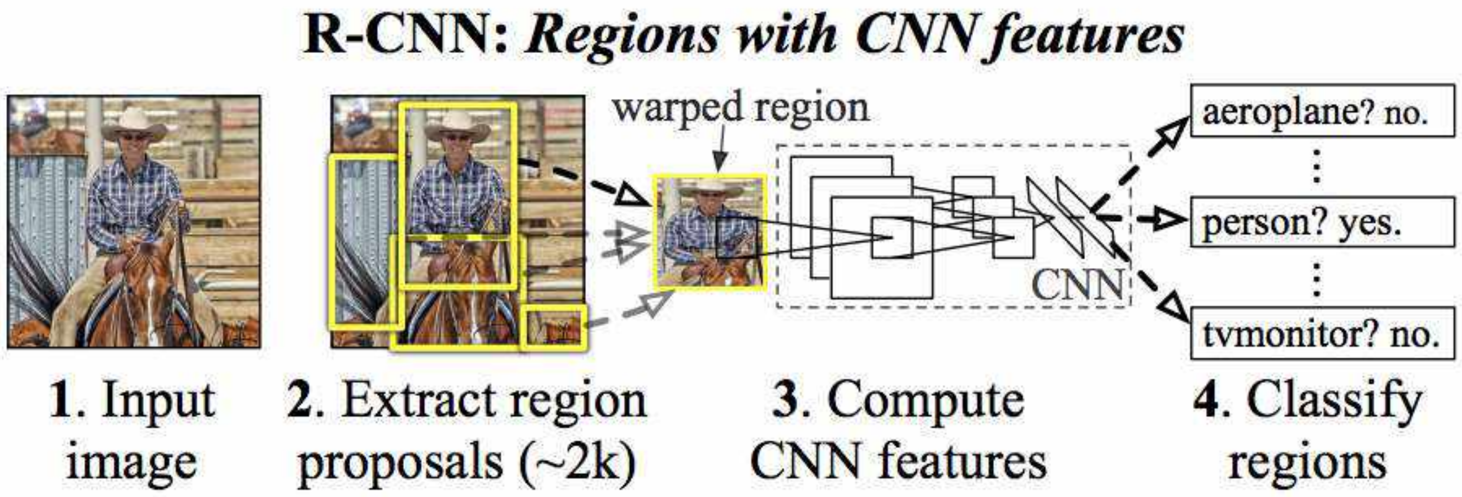




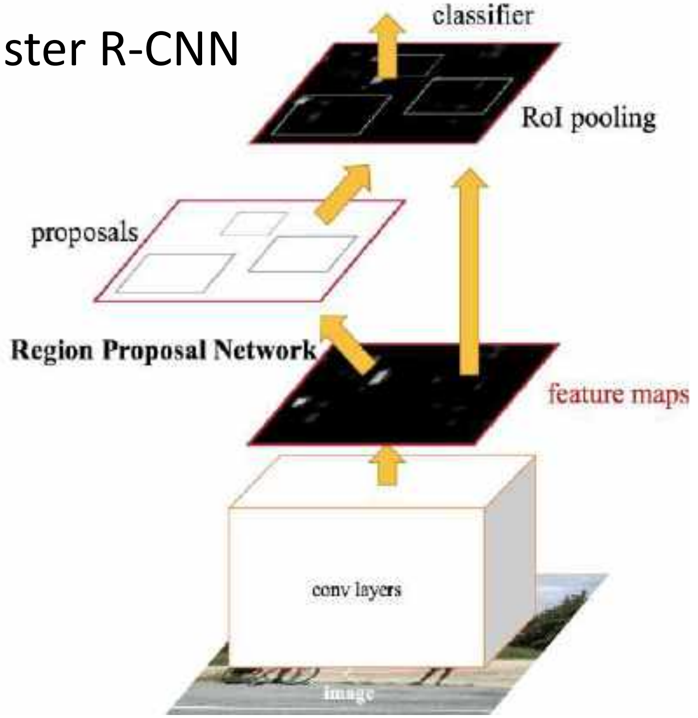
$$f(x, y) = \sum_i p_b(i) - \sum_i p_w(i)$$

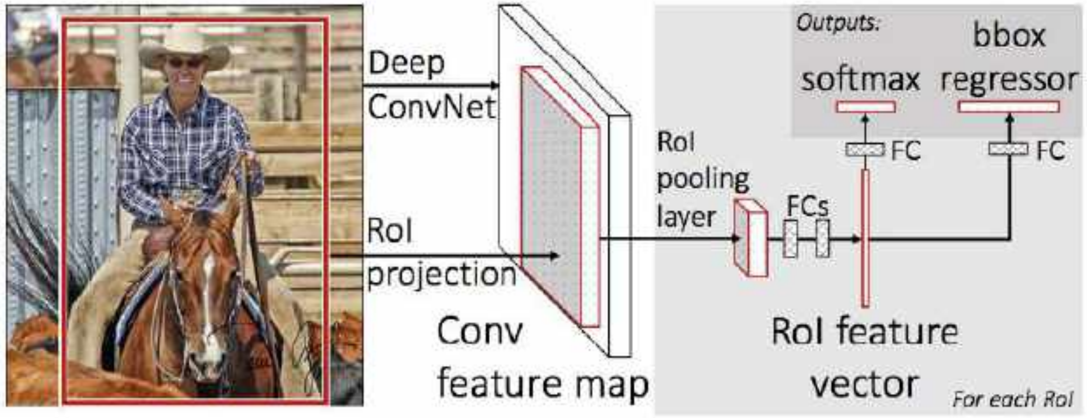
- P. Viola and M. Jones. *Rapid Object Detection using a Boosted Cascade of Simple Features.* CVPR 2001.
- P. Viola and M. Jones. *Robust Real-Time Face Detection.* IJCV 57(2), 2004.

Approach 2: Generate a few Region Proposals

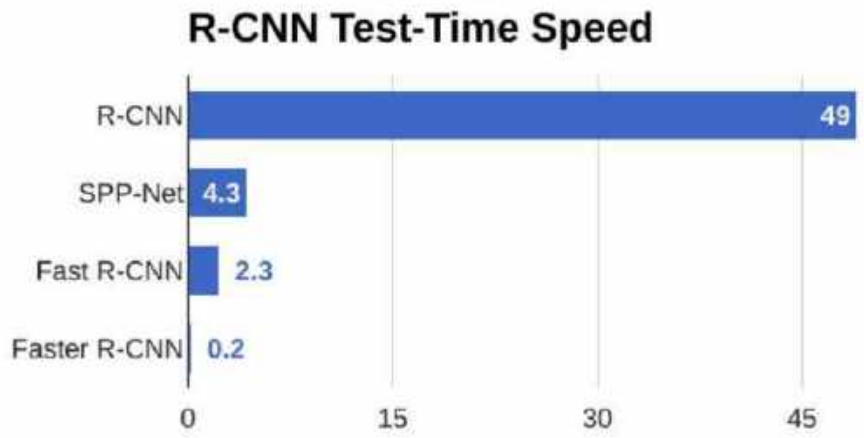


Faster R-CNN



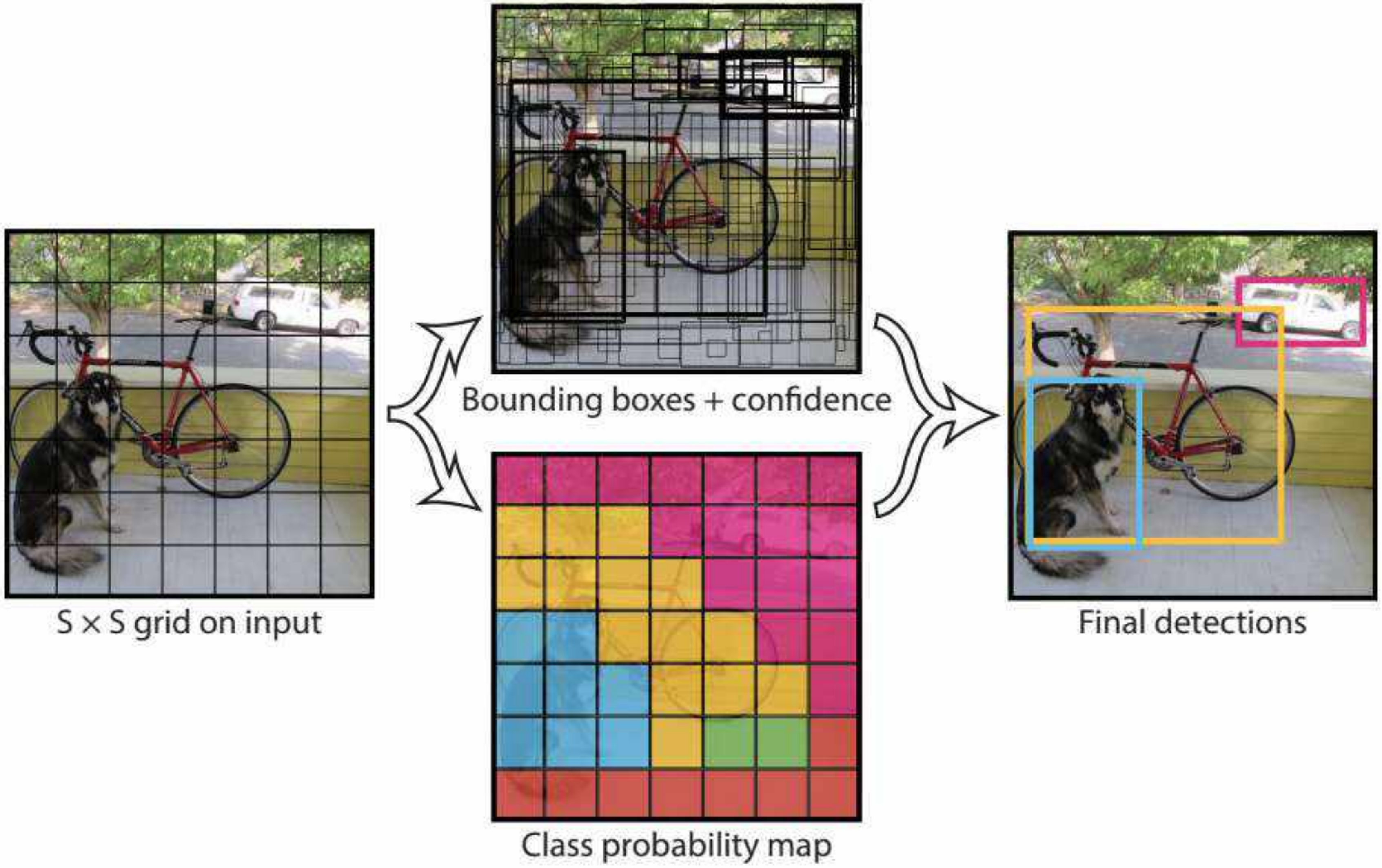


Fast R-CNN

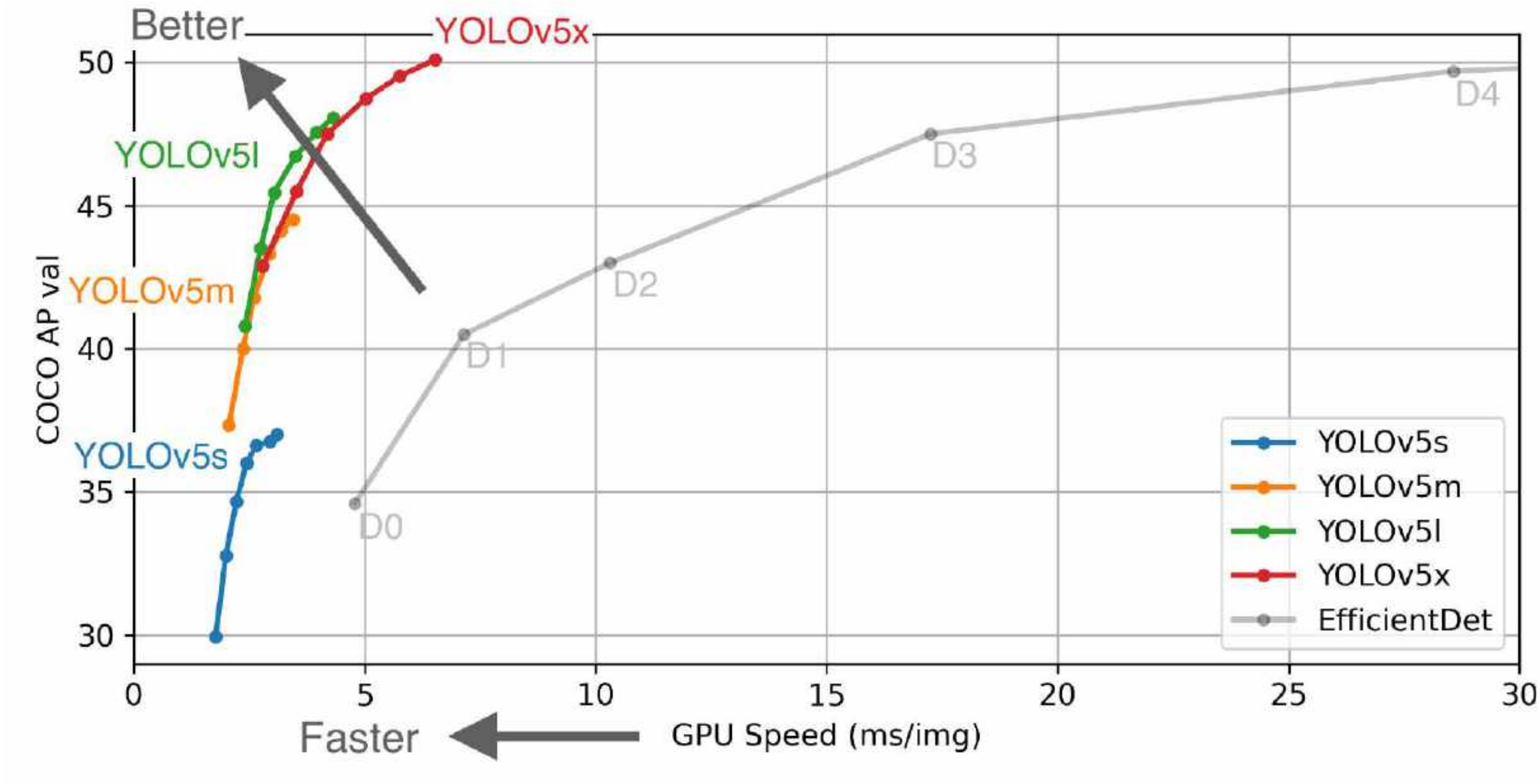


Approach 3: Single Stage for Speed

- YOLO v1 .. v5
- SSD
- RetinaNet
- Efficient-Det



Accuracy vs. Speed of Detection



Further Reading

- Most modern approaches is a potpourri of different Network Architectures, Additional Connections, Normalizations, Activations, Data Augmentation, Regularization, Spatial Attention, Training strategies, etc.
- Most popular Dataset: MS-COCO
- Nice Summary of Approaches:
  – Alexey Bochkovskiy, Chien-Yao Wang, Hong-Yuan Mark Liao, “YOLOv4: Optimal Speed and Accuracy of Object Detection”

A few other topics in Computer Vision

- Photometry & Radiometry
  - Model Surface Properties
  - Measurement, Inspection, …
- Computational Imaging
  - Computational Illumination
  - Computational Optics
  - Modified/Simplified Sensors
  - Post Processing
- Shape-From-X
  - Multi-View
  - Single Image
  - Other Cues
- Video Processing
  - Video Prediction
  - Behaviour Classification
- Image Generation
  - Graphics + Vision
  - GANs
- Application Areas
  - Biometrics
  - Robotics
  - Medical Image Analysis
  - …

Thanks!!

Questions?