THE UNIVERSITY OF MELBOURNE

Depth and stereo

Semester 2, 2021

Kris Ehinger + Tom Drummond

Ames Room

https://www.youtube.com/watch?v=Ttd0YjXF0no

Outline

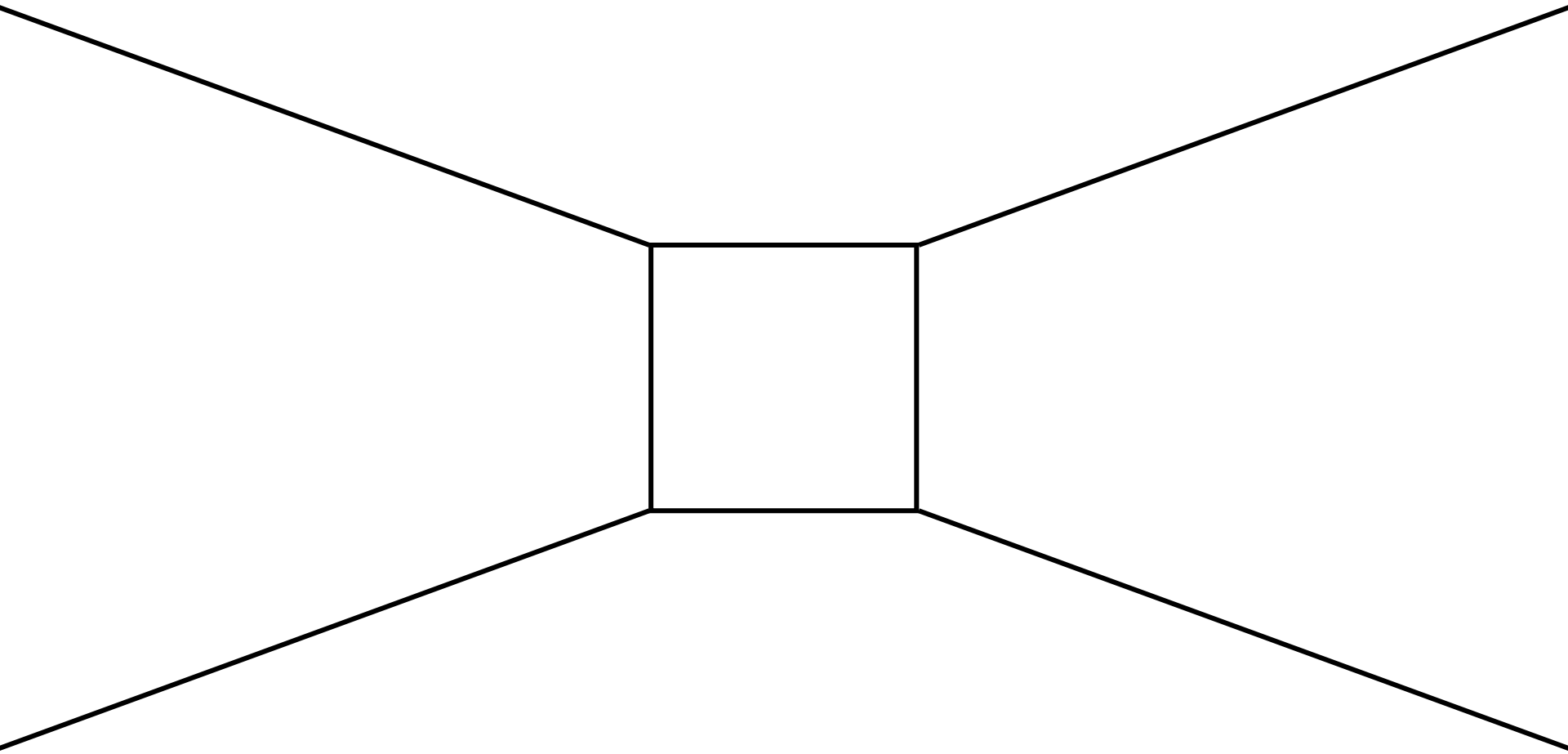
- Cues for depth
- Stereo
- Single-view depth

Learning outcomes

- Explain the cues available for depth perception in single-view and multi-view images
- Compute depth from disparity
- Explain how a standard stereo depth algorithm works
- Explain machine-learning approaches to single-view depth

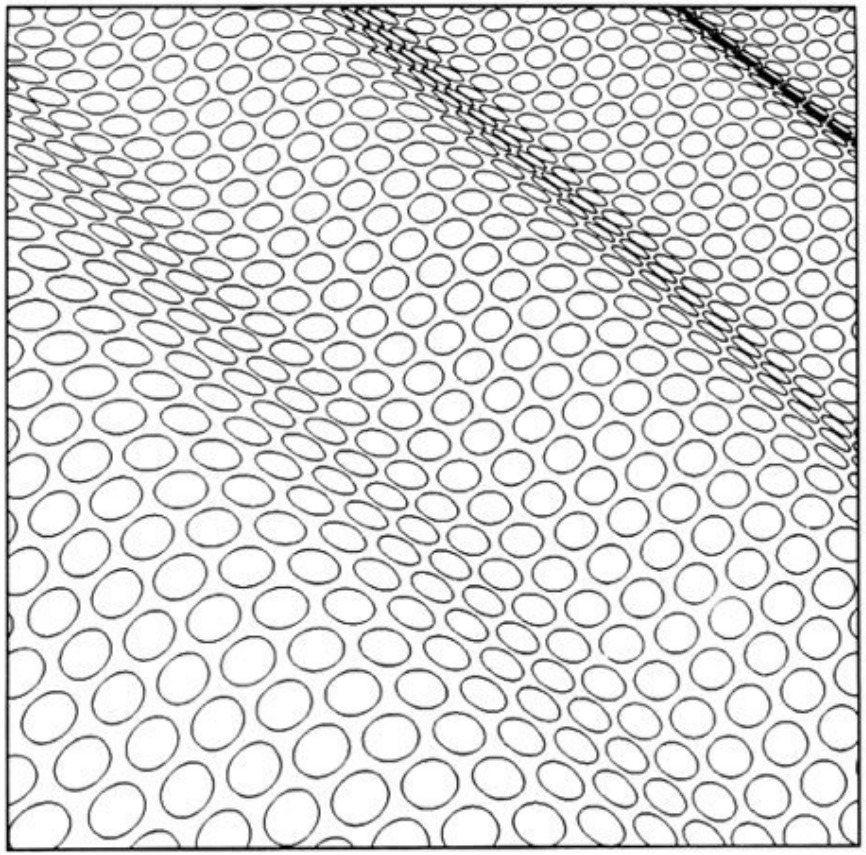
Cues for depth

Perspective



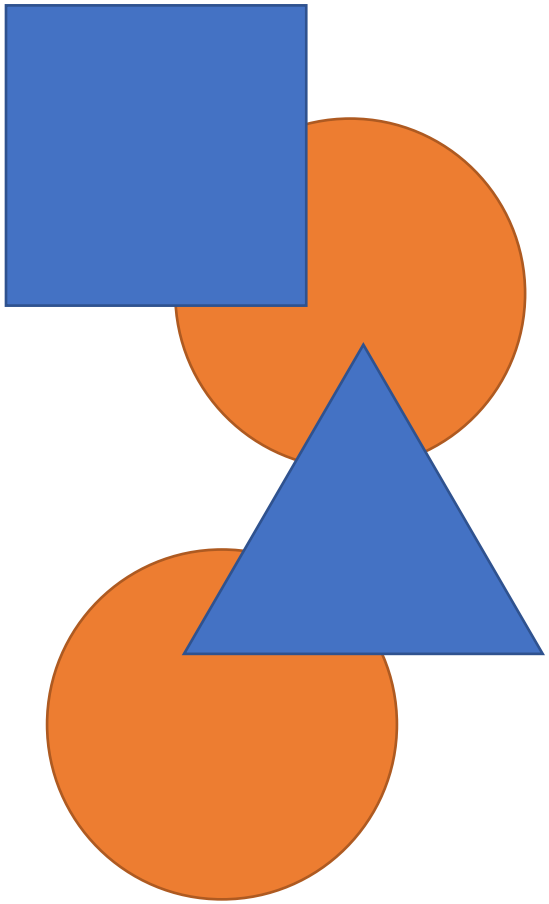
Perspective

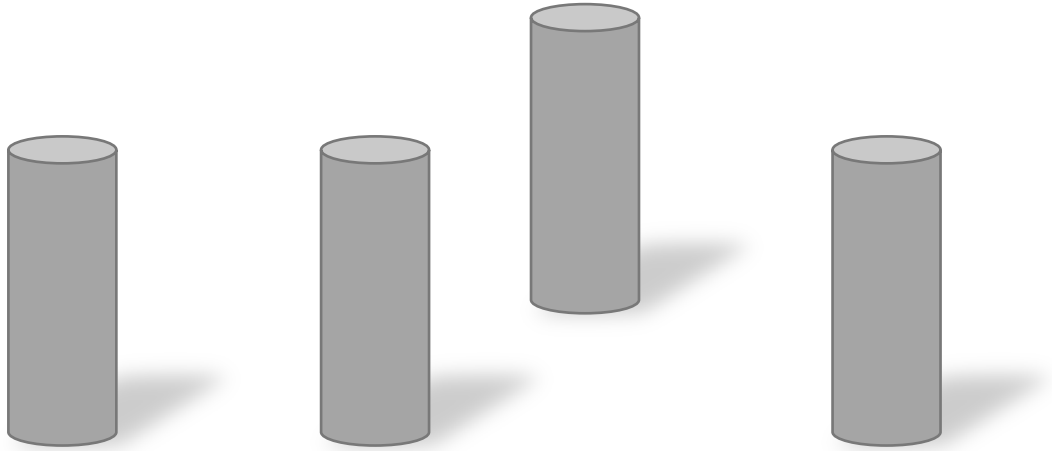
Texture



https://www.beautifullife.info/urban-design/10-crazy-3d-optical-illusions-will-blow-mind/
http://www.cns.nyu.edu/~david/courses/perception/lecturenotes/depth/depth-size.html

Object position and occlusion





Object knowledge

https://en.wikipedia.org/wiki/Forced_perspective, http://viperlibnew.york.ac.uk/

Binocular stereo

Image: The National Archives (United Kingdom)

Motion parallax

https://students.unimelb.edu.au/student-precinct/project-updates/drones-eye-view2

Cues for depth

- 2D images contain a variety of information for depth perception
- Cues available in a single view include perspective, texture, and object cues
- More accurate depth information can be obtained by combining multiple views (stereo, motion)

Stereo

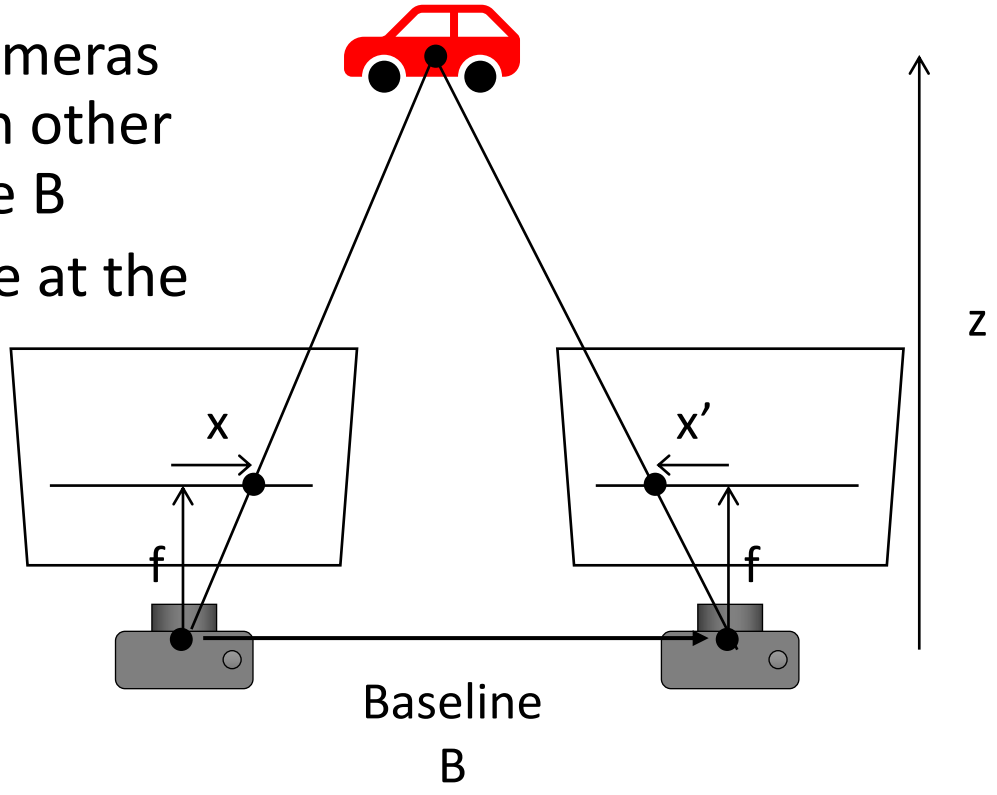
Depth from stereo

- Stereo pair: images from two cameras with a horizontal shift in camera position

Image: https://www.instructables.com/3D-Stereoscopic-Photography/

Depth from stereo

- Assume:
  - Image planes of cameras are parallel to each other and to the baseline B
  - Camera centres are at the same height
  - Focal lengths f are the same
- Goal: find z



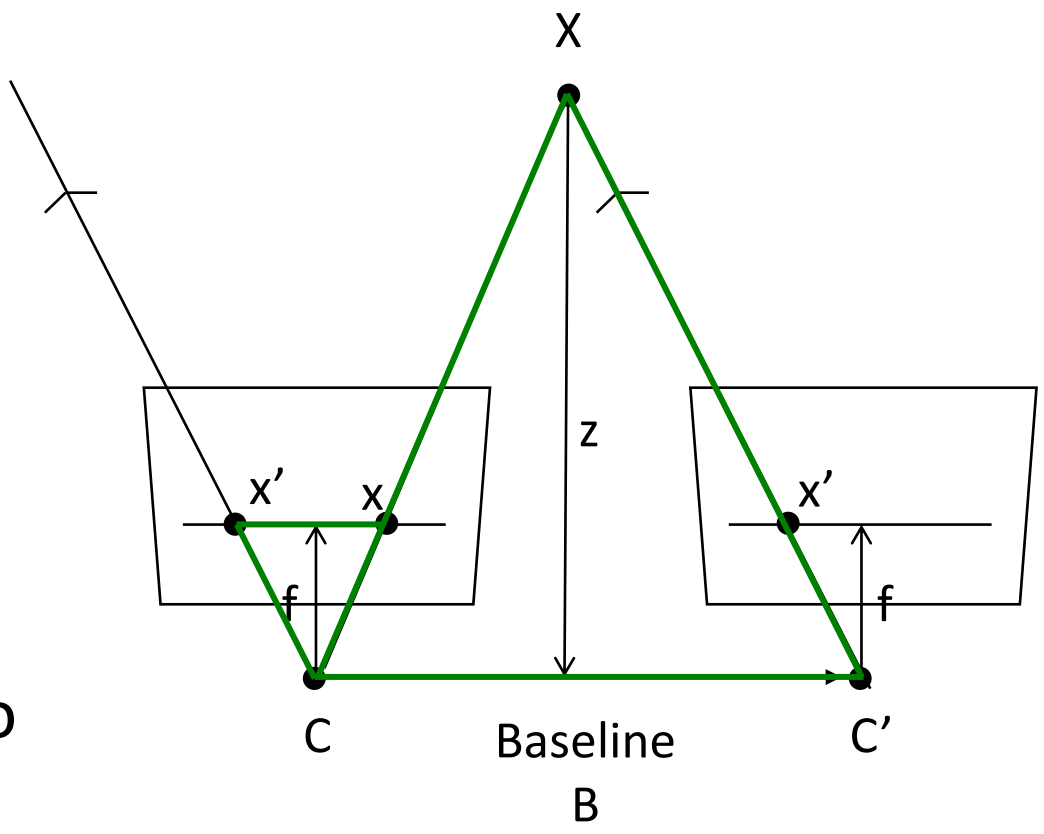
Depth from stereo

- Solve for z:

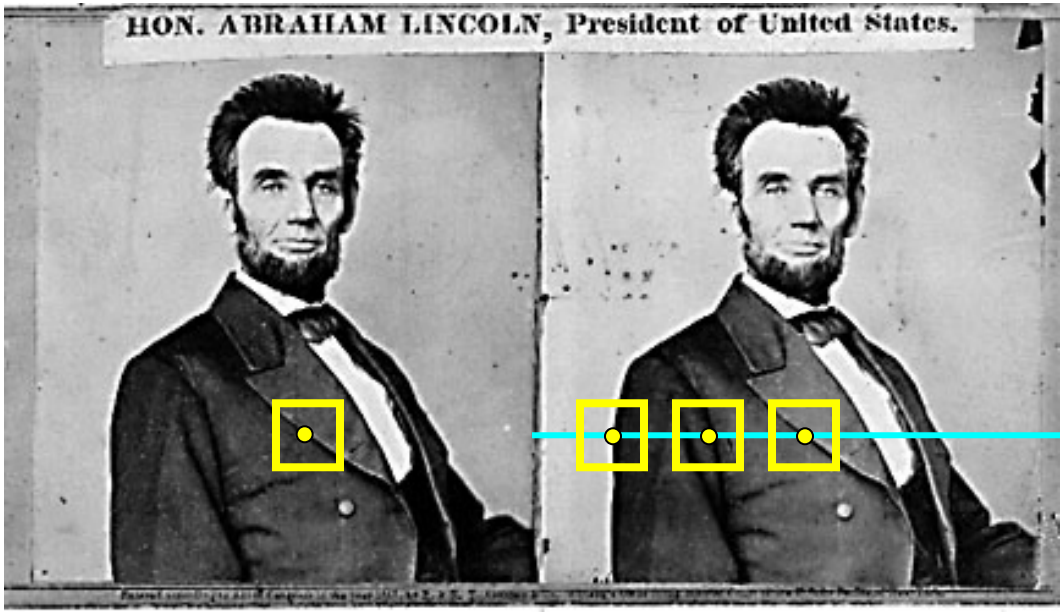
$$\frac{z}{B} = \frac{f}{x - x'}$$

$$z = \frac{fB}{x - x'}$$

- Distance z is inversely proportional to disparity x-x’



Basic stereo matching algorithm

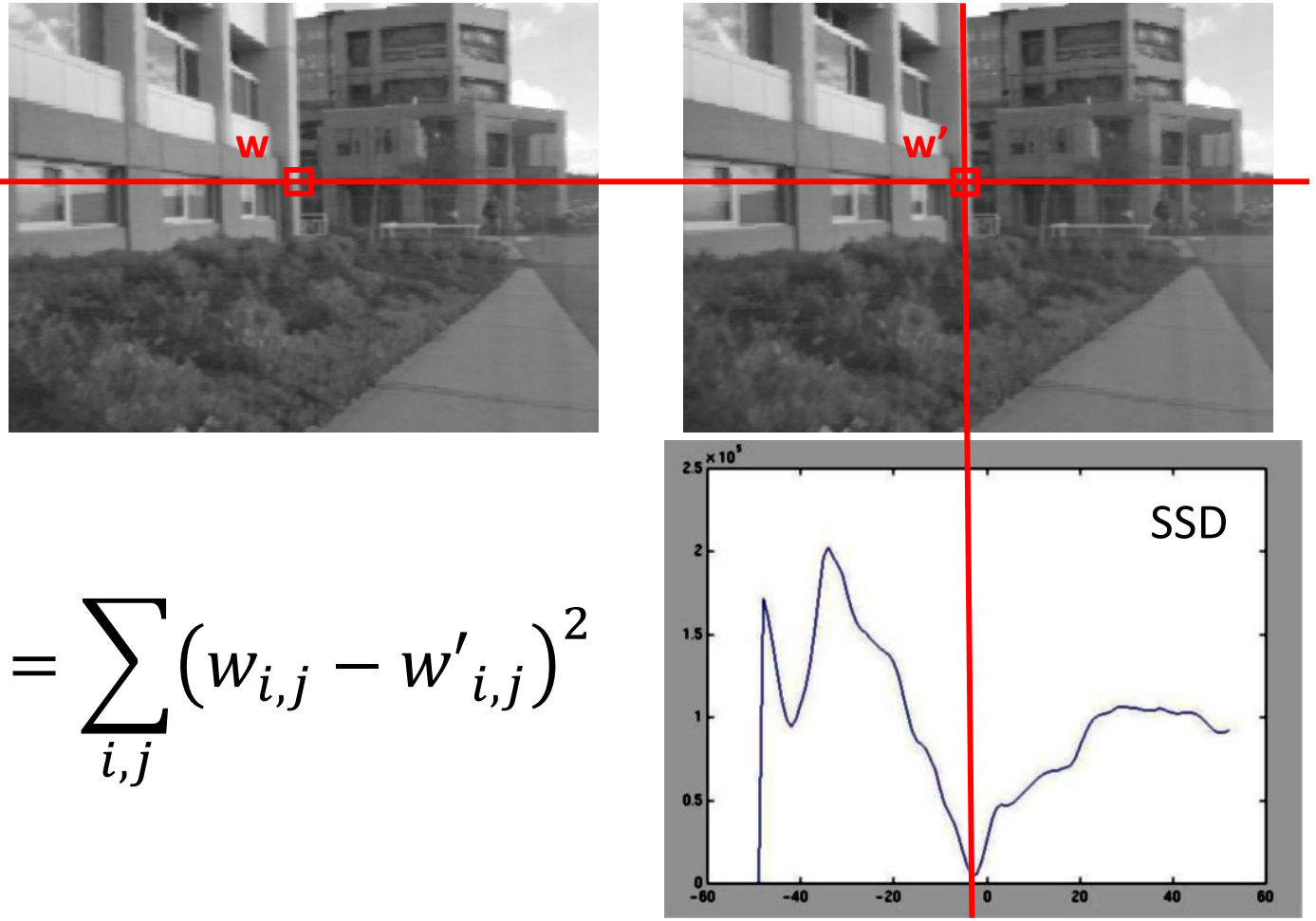


- For each pixel x in one image
  - Scan a horizontal line in the other image, find best match x'
  - Compute disparity x-x' and compute depth = (fB)/(x-x')

Image: http://www.johnsonshawmuseum.org/

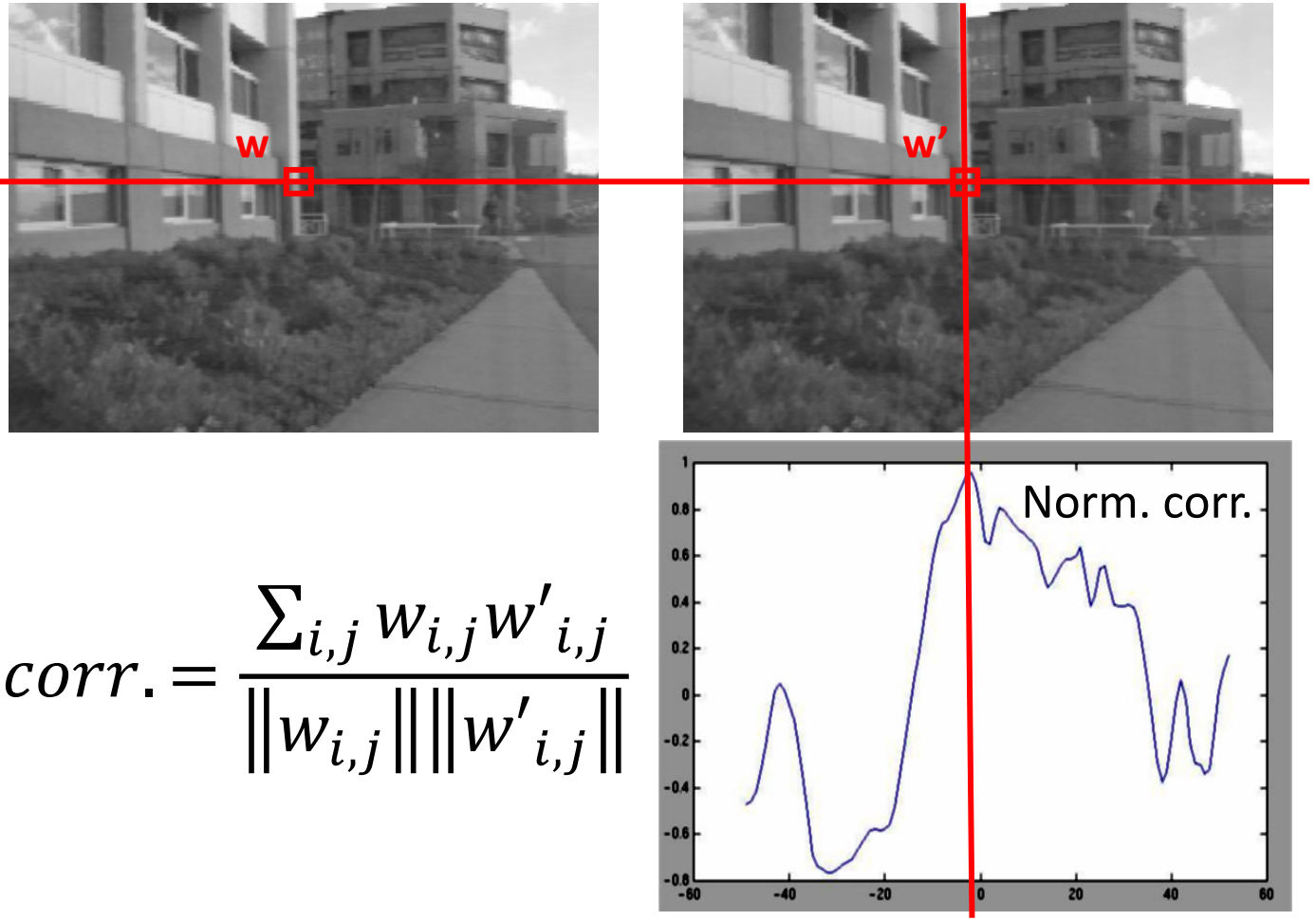
Basic stereo matching

Scanline



$$SSD = \sum_{i,j} \left(w_{i,j} - w'_{i,j} \right)^2$$

Basic stereo matching



$$norm.corr. = \frac{\sum_{i,j} w_{i,j} w'_{i,j}}{\|w_{i,j}\| \|w'_{i,j}\|}$$

Example: Autostereogram

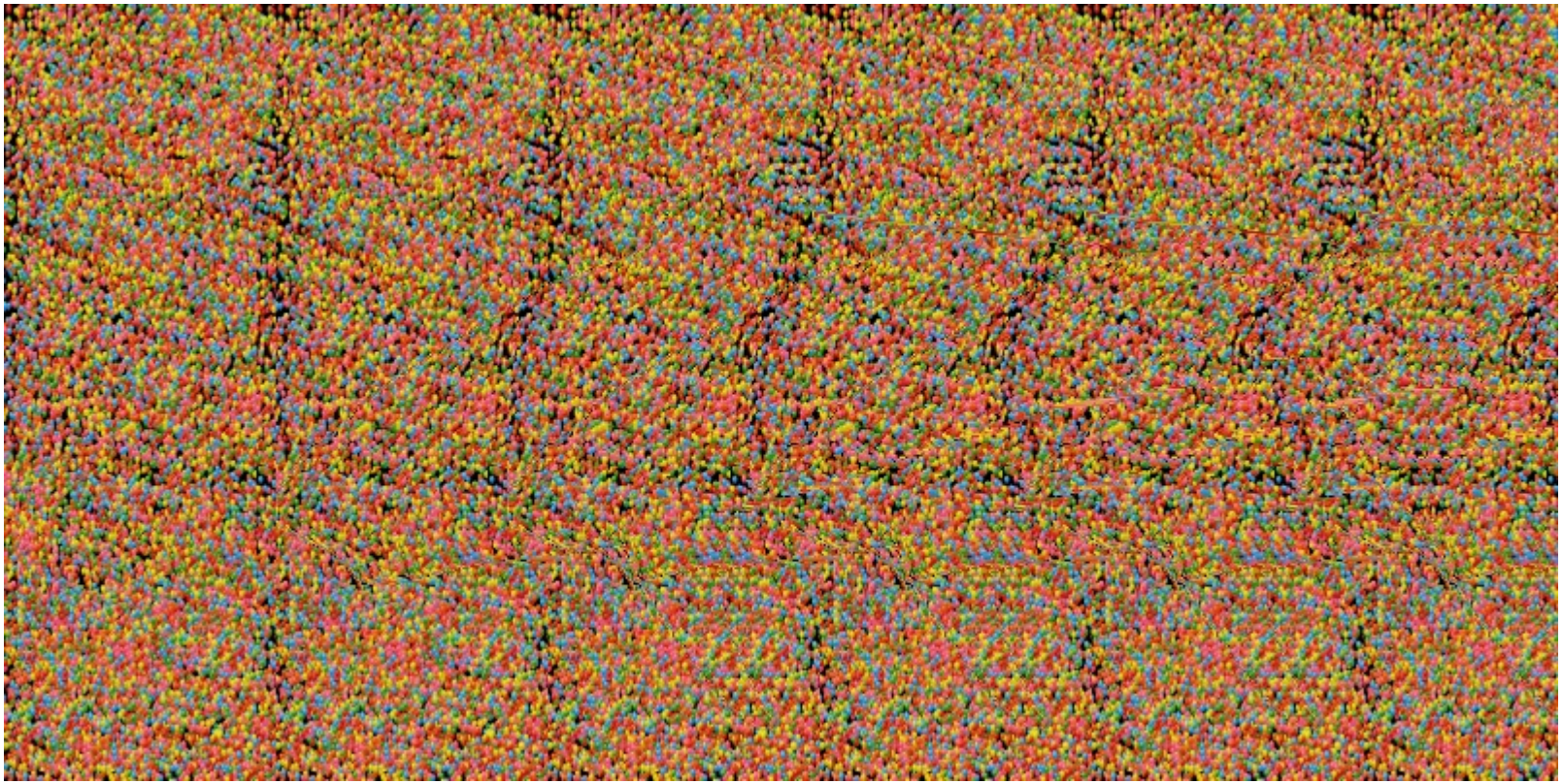
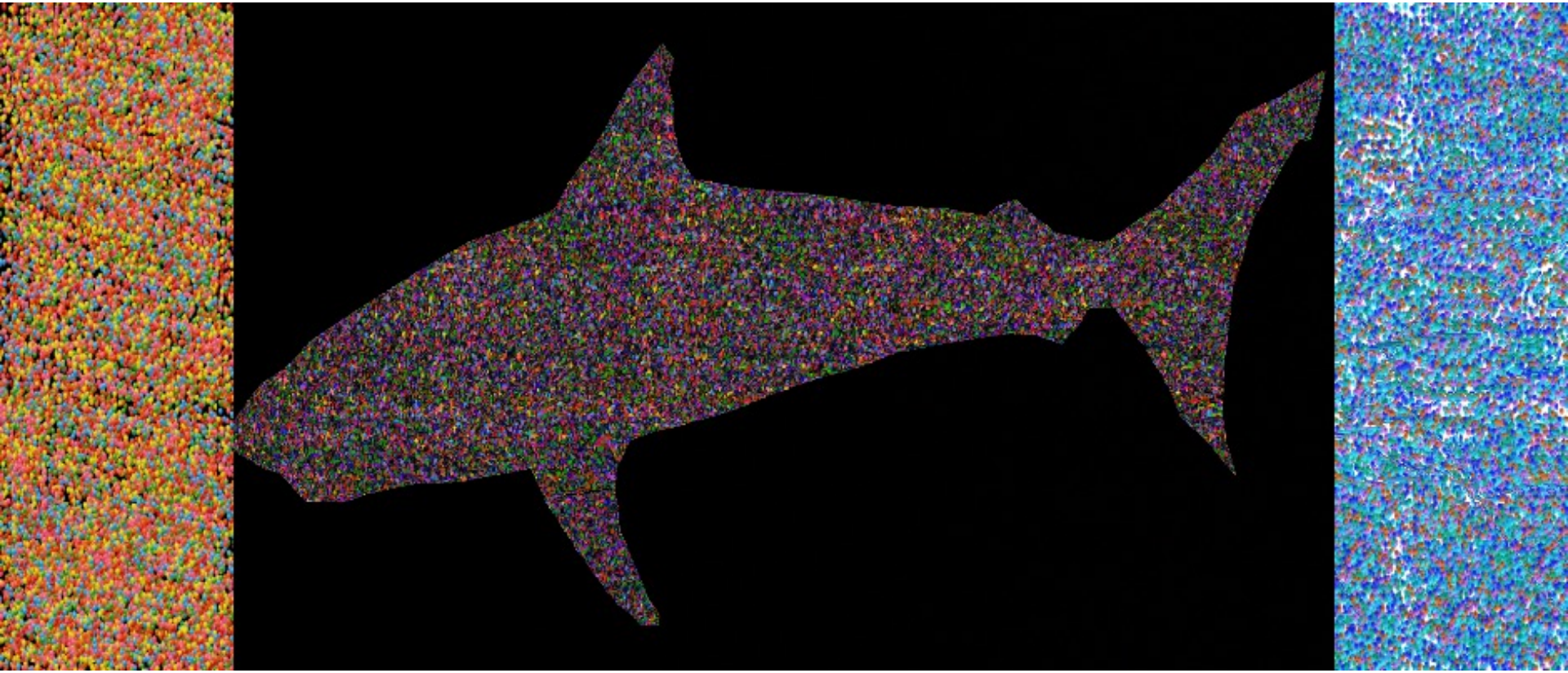


Image: Fred Hsu (2005)

Example: Autostereogram

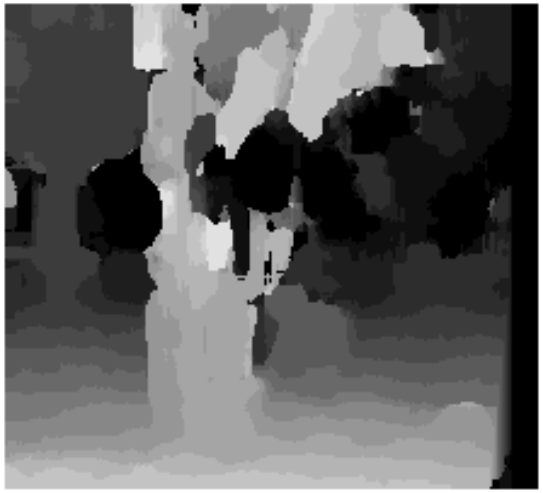


(image) – (offset image)

Effect of window size

Depth map

Window size = 3x3 px



Window size = 20x20 px

- Smaller window = finer detail, but more noise
- Larger window = smoother depth, but lacking detail

Image: S. Seitz

Match failures

Image: http://www.johnsonshawmuseum.org/

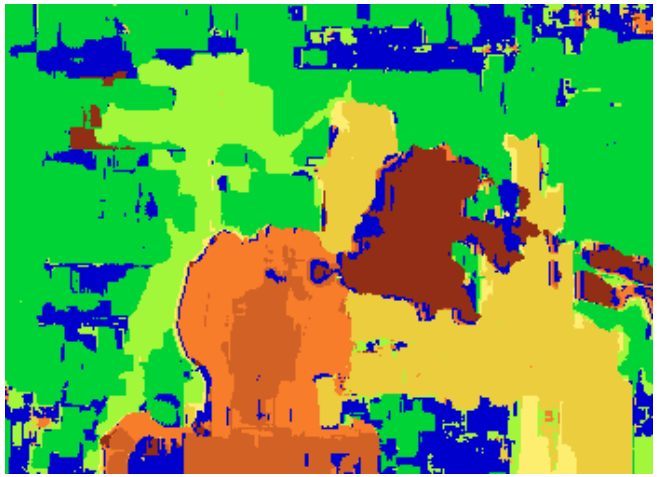
Match failures

Reflective object

Painted object

Image: Muryy, Fleming, & Welchman (2014)

Window matching result



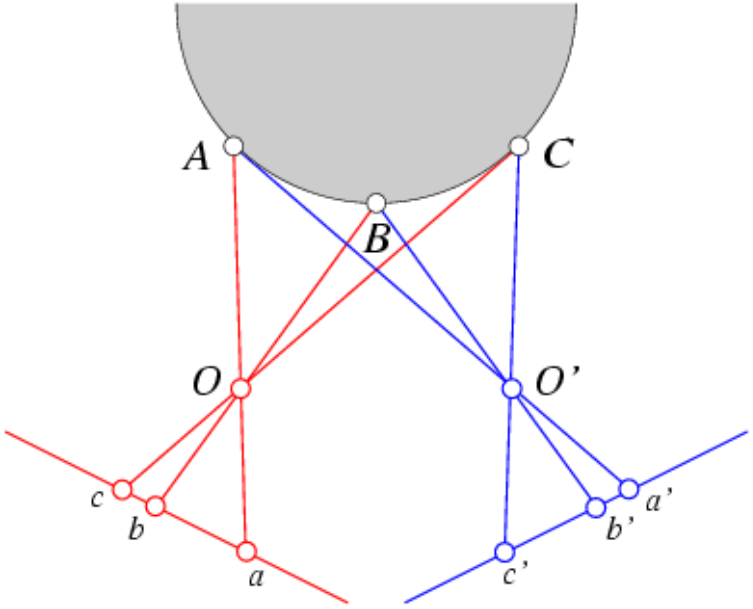
Estimated depth

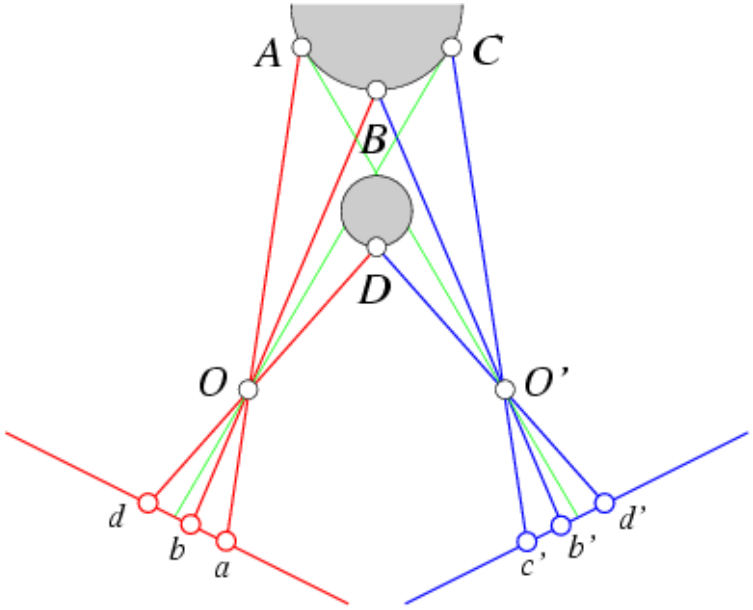
Image: S. Seitz

Additional constraints?

- Individual matches are often ambiguous
- However, the set of matches should obey additional constraints:
  - Uniqueness: a point in one view has no more than one match in the other view
  - Ordering: corresponding points should be in the same order in both views
  - Smoothness: disparity values should change smoothly (for the most part)

Ordering constraint

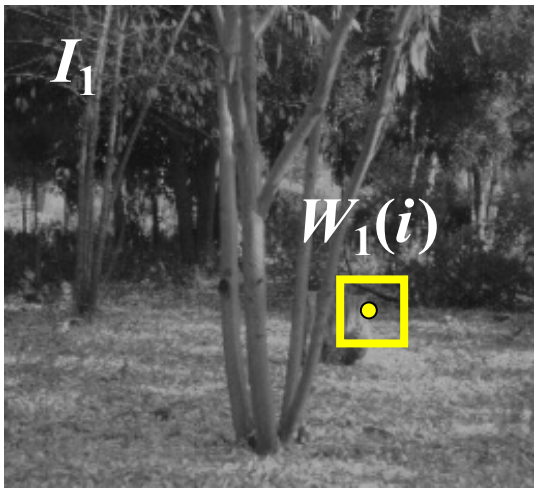


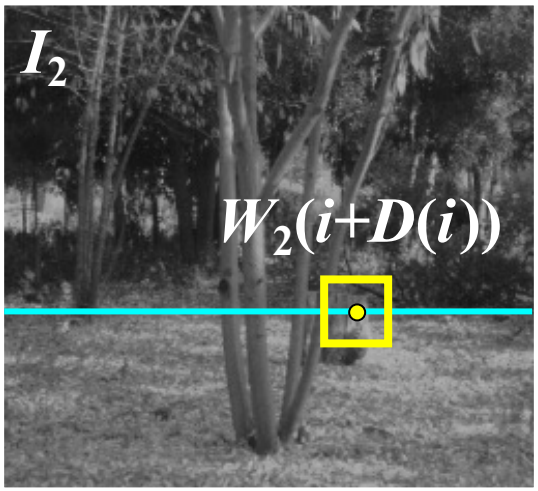


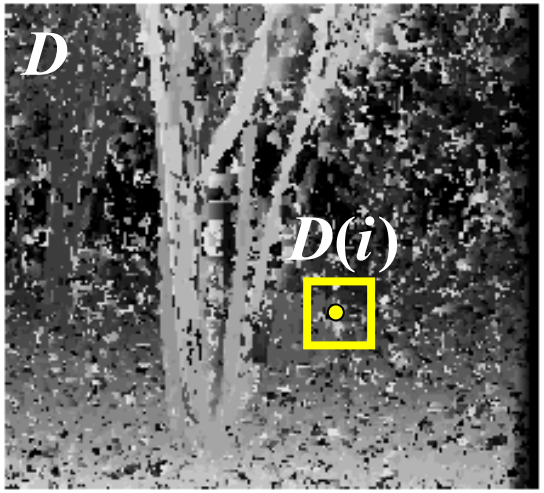
Ordering constraint does not always hold

Image: S. Seitz

Applying constraints







$$E(D) = \sum_i \left(W_1(i) - W_2(i + D(i)) \right)^2 + \lambda \sum_{\text{neighbors } i,j} \rho\left(D(i) - D(j) \right)$$

Match quality

Smoothness

Minimize $E(D)$ using an optimisation method such as graph cuts (Boykov, Veksler, & Zabih, 2001)

Image: S. Seitz

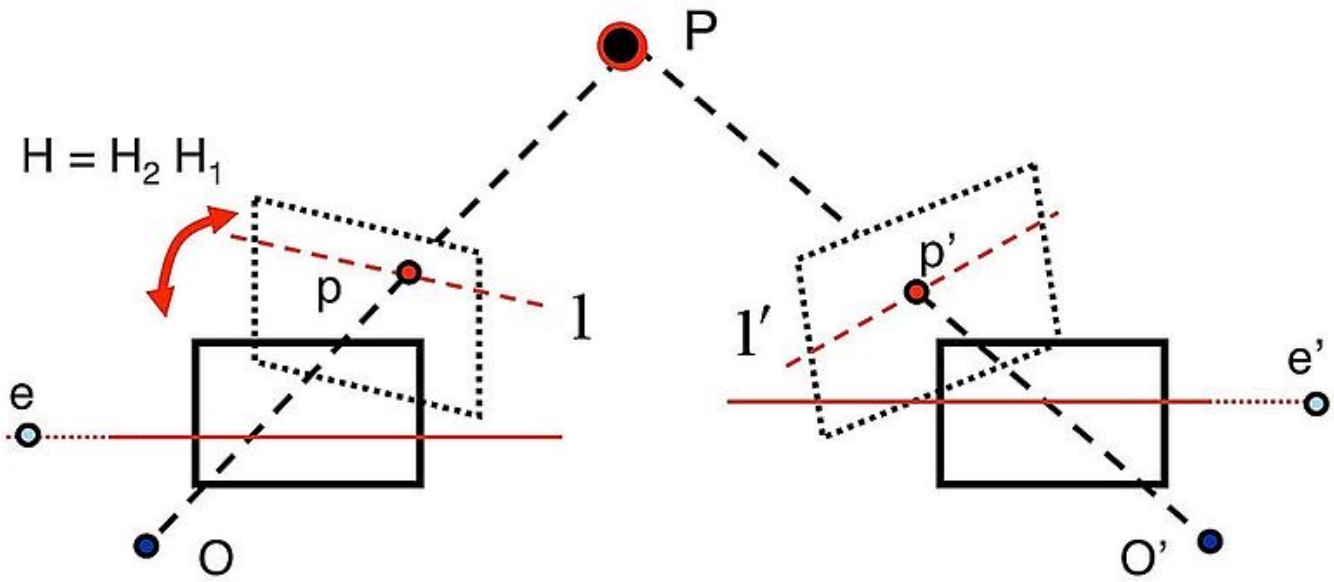
Window matching result

Estimated depth from graph cuts

Image: S. Seitz

Rectification

- What if the image planes are not parallel?
- **Rectification** finds projective transform that maps each image to the same plane



Summary

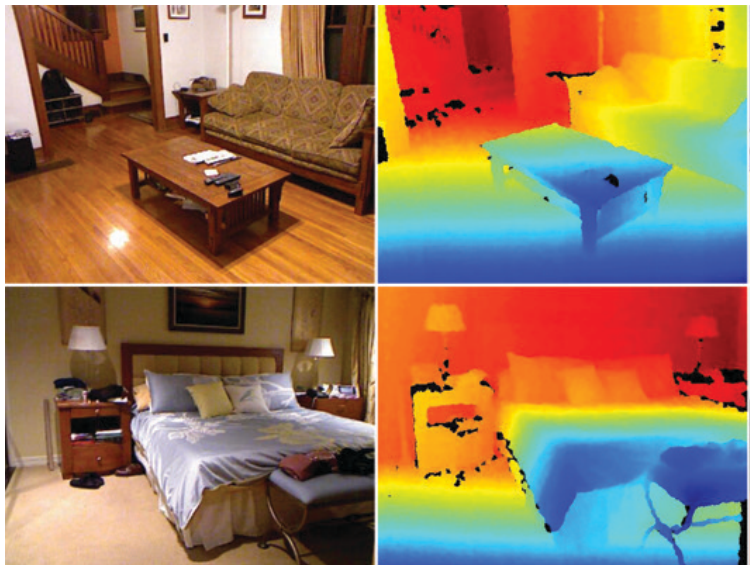
- Stereo depth is computed from disparity (difference in position of points in two views of a scene)
- Difficult to solve in practice because individual point matches are ambiguous
- Solution generally involves additional constraints (e.g., smoothness)

Single-view depth

Depth from single images

Supervised depth classification

- Treat depth estimation as a classification task: For each pixel in image, predict distance from camera
- Train on images with annotated depth maps



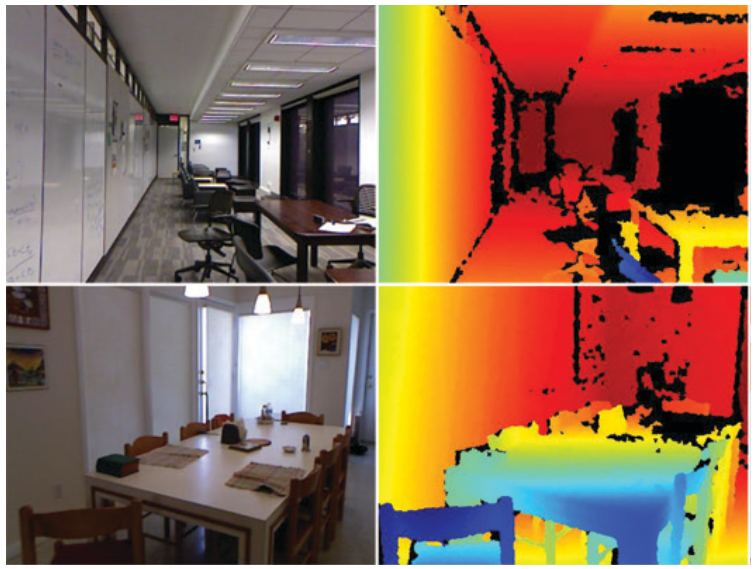
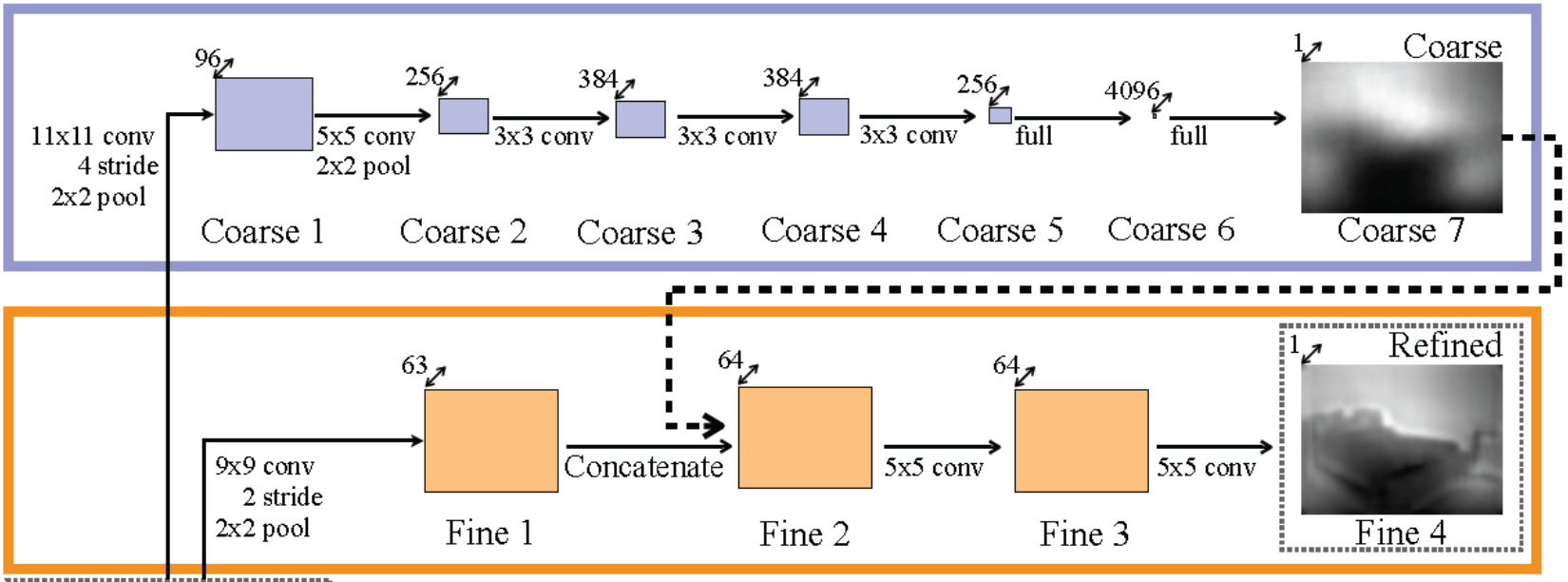


Image: NYU Depth Dataset V2

Supervised depth classification



		Coarse					Fine
Layer	input	1	2,3,4	5	6	7	1,2,3,4
Size (NYUDepth)	304x228	37x27	18x13	8x6	1x1	74x55	74x55
Size (KITTI)	576x172	71x20	35x9	17x4	1x1	142x27	142x27
Ratio to input	/1	/8	/16	/32	–	/4	/4

Eigen, Puhrsch, Fergus (2014)

Loss function?

- Images may have a very large range of depths – a loss based on log(depth) may work better than absolute depth
- Mean depth of scenes can vary widely (e.g., closet vs. stadium). To discourage models from simply learning mean depth, consider scaling the loss function so that it is similar for different scenes.

Supervised classification results

Image | Coarse result | Final result | Ground truth depth

Eigen, Puhrsch, Fergus (2014)

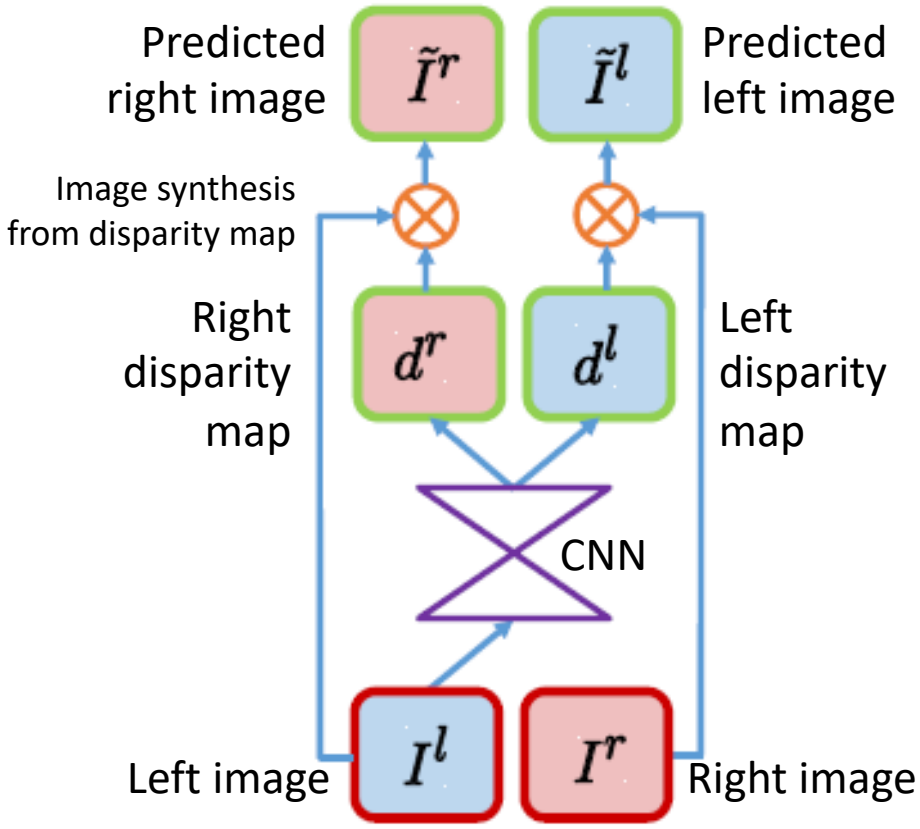
Depth from disparity

- Instead of training on annotated depth maps, train on stereo image pairs
- Advantage: stereo image pairs can be produced with standard cameras, while depth maps require special equipment (e.g., LiDAR)

Depth from disparity

- Input: one image from a stereo pair (e.g., left)
- Learn to predict the disparity map that will produce the other image (e.g., right)
- Distance from camera to surfaces can be computed from the disparity map
- Train on stereo pairs

Depth from disparity



- Loss is sum of:
  - Appearance loss (difference between original and predicted image)
  - Disparity smoothness loss
  - Left-right consistency loss (difference between disparity maps)

Godard, Aodha, & Brostow (2017)

Depth from disparity results

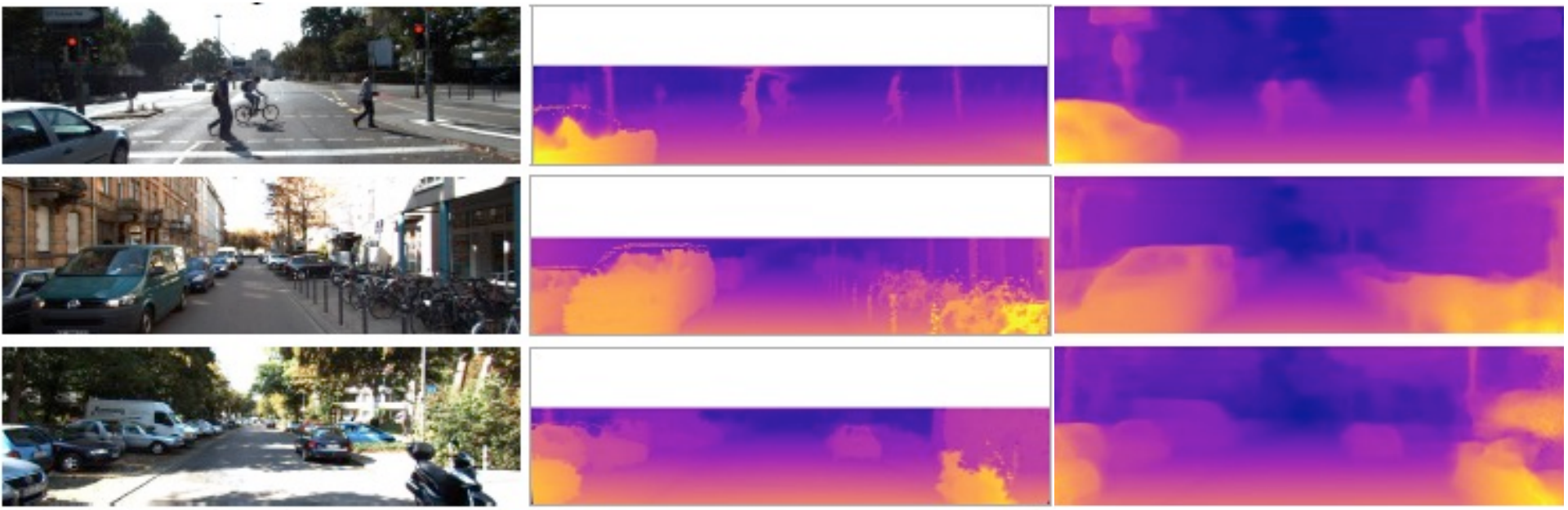


Image | Ground truth depth | Estimated depth

Godard, Aodha, & Borstow (2017)

Summary

- Machine learning methods can be trained to estimate depth from single images
- Advantages:
  - Does not require multiple views / stereo camera
- Disadvantages:
  - “Blurry” depth at object edges (but can be combined with edge maps for better results)
  - Models may not generalise well to new contexts

Summary

- 2D images provide multiple cues for 3D depth (single-image and multi-image cues)
- It is possible to compute depth from single images, but more accurate depth measurements can be obtained from multiple views (e.g., stereo)