Stereo

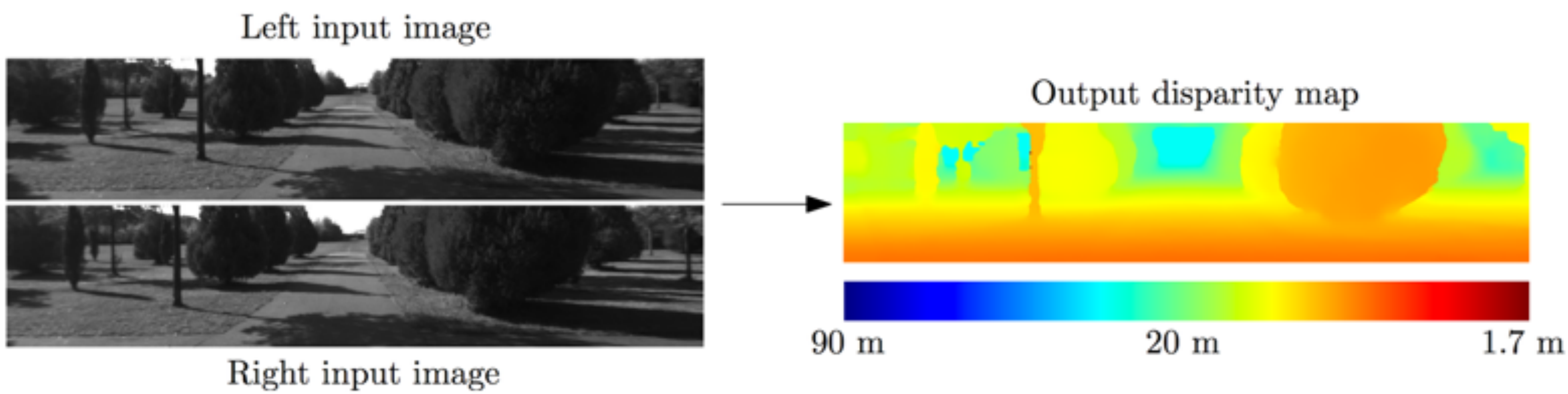
Wenjie Luo
CSC2541

Feb 2nd, 2016

Outline

- Problem specifics
- Matching, conv nets
- Smoothing(CRF), post-processing
- Discussion

Driving a car



Source: Zbontar & LeCun

- Understanding surrounding area: depth
- Depth is crucial for making certain decisions

Why depth

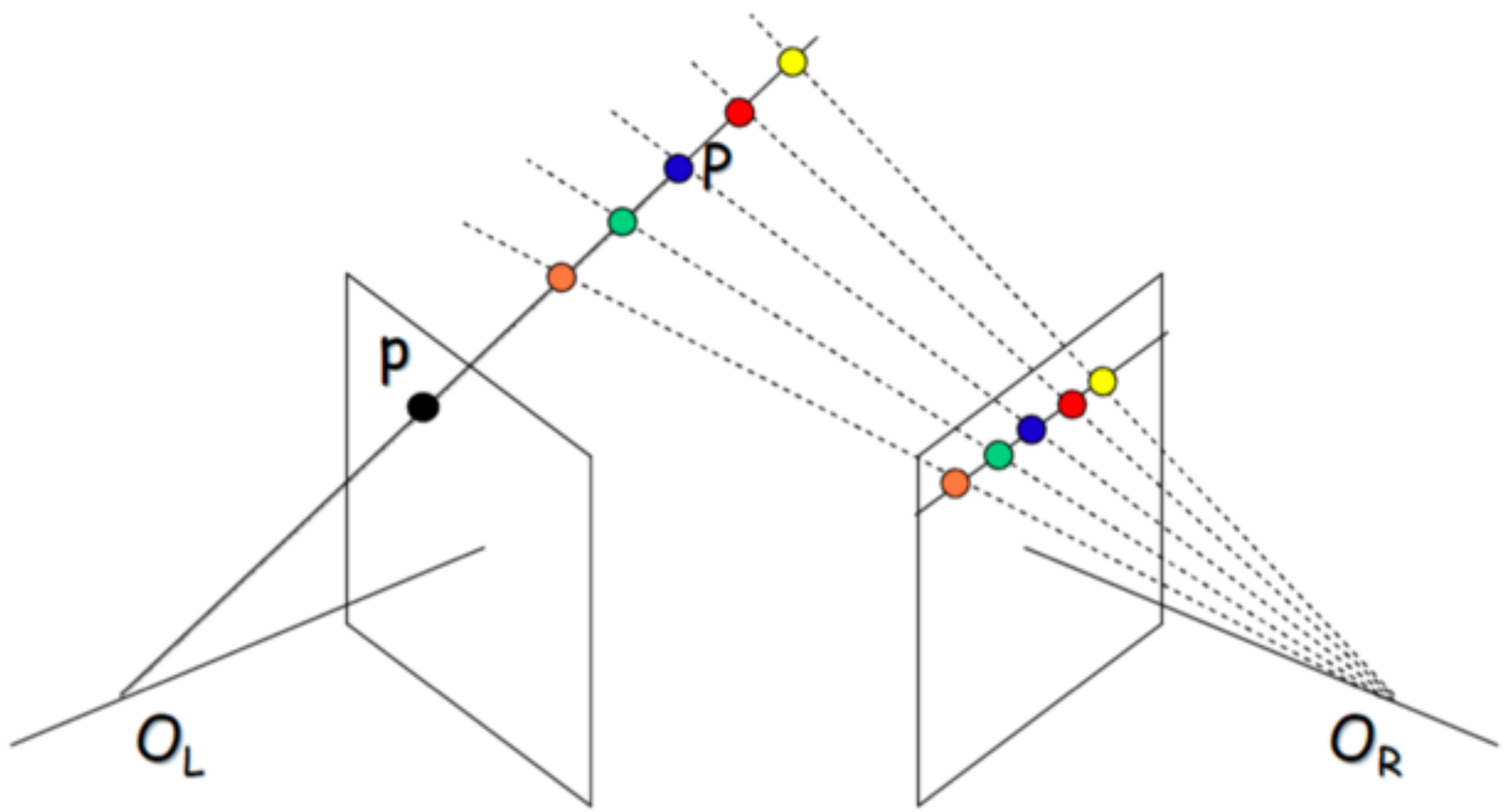
Source: L. Lazebnik

- Why it's difficult
    - Ambiguous, correspondence, occlusion..

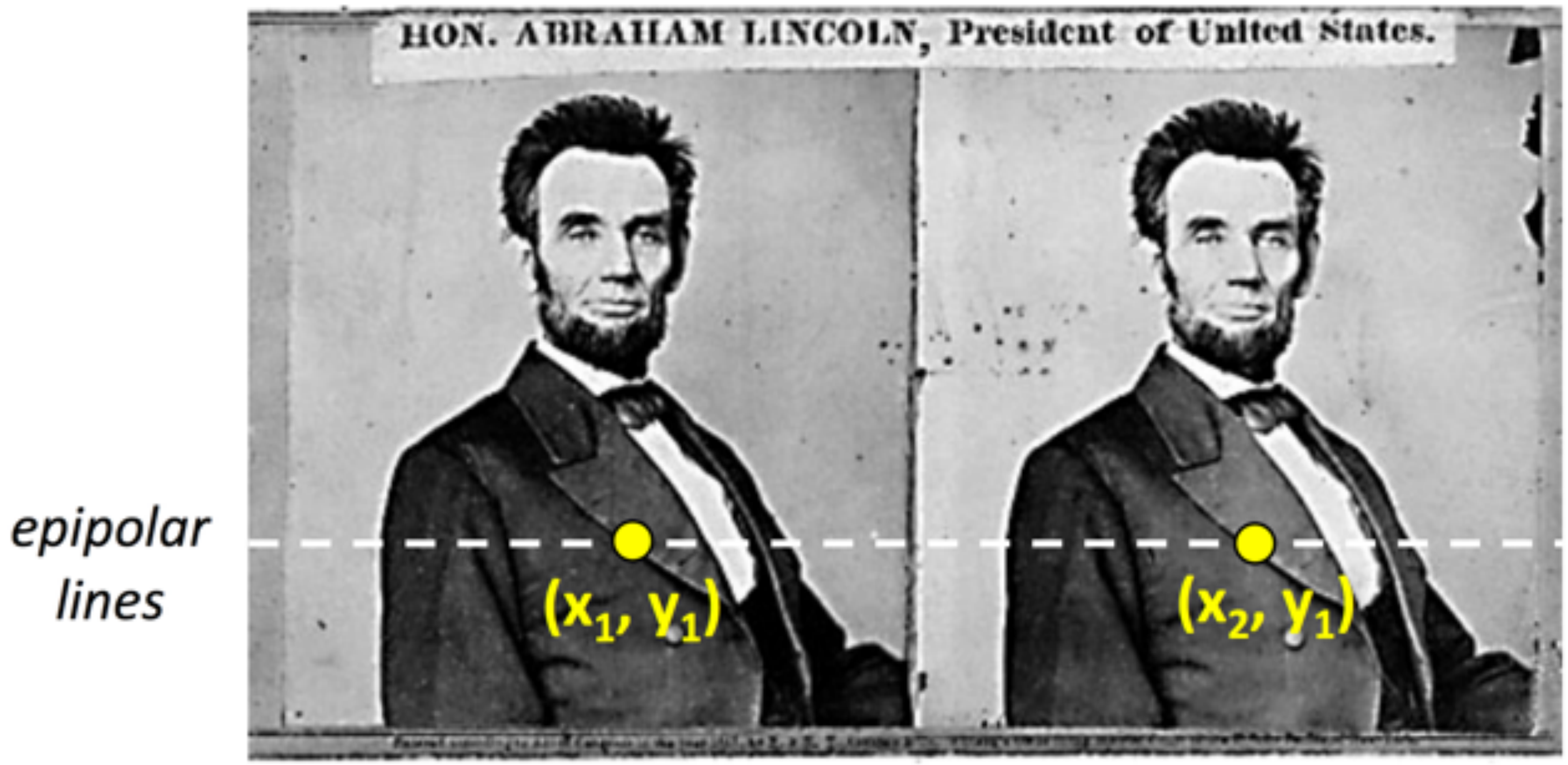
- How to get depth
    - Perspective, relative size, occlusion, texture gradients
    - Single image, *stereo*, multiple-view

Stereo

- Estimate depth from stereo images.



Source: R. Urtasun



Two images captured by a purely horizontal translating camera
(*rectified* stereo pair)

- Depth is inversely proportional to disparity.

$$Z = f\frac{B}{d}$$

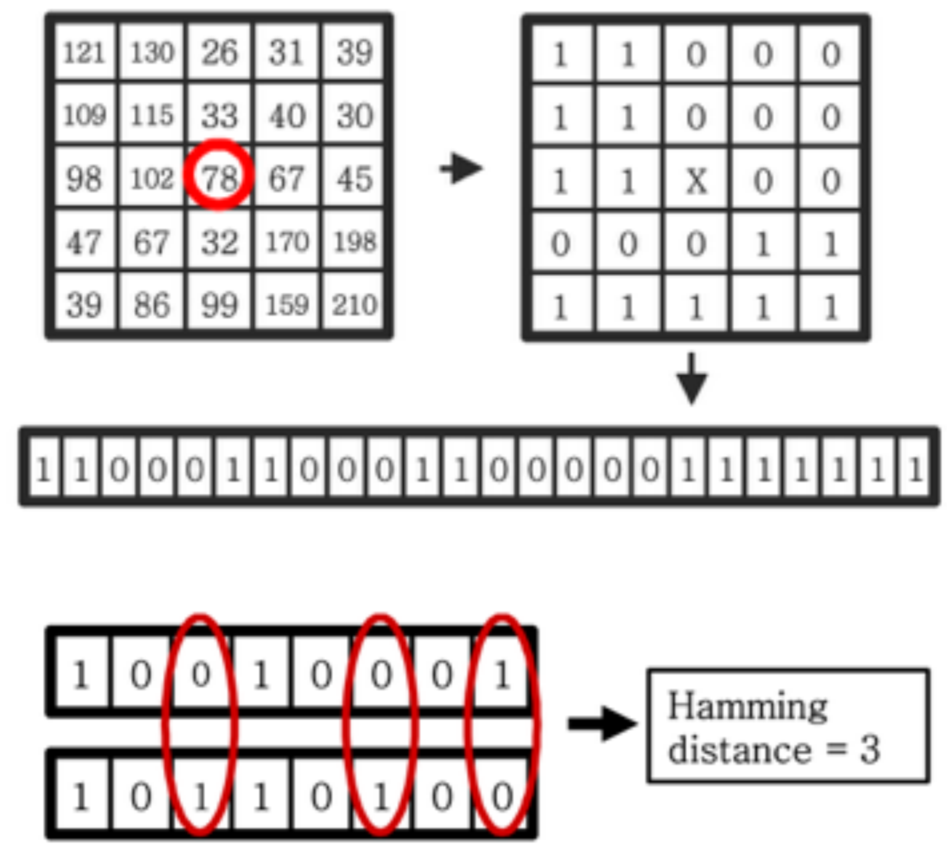
Z: depth; f: focal length; B: baseline; d: disparity

We need..

- Info on camera pose(Calibration)
  - Fixed and known
- Correspondances on image locations(Matching)
  - Hand-crafted feature
  - Learnable feature from Conv-Nets
- Refinement in practice
  - Smoothing

Fixed feature

- Image intensity, color
- Image gradient
- Census transform
    - local spatial structure
    - hamming distance



Source: Young Baik et.al.

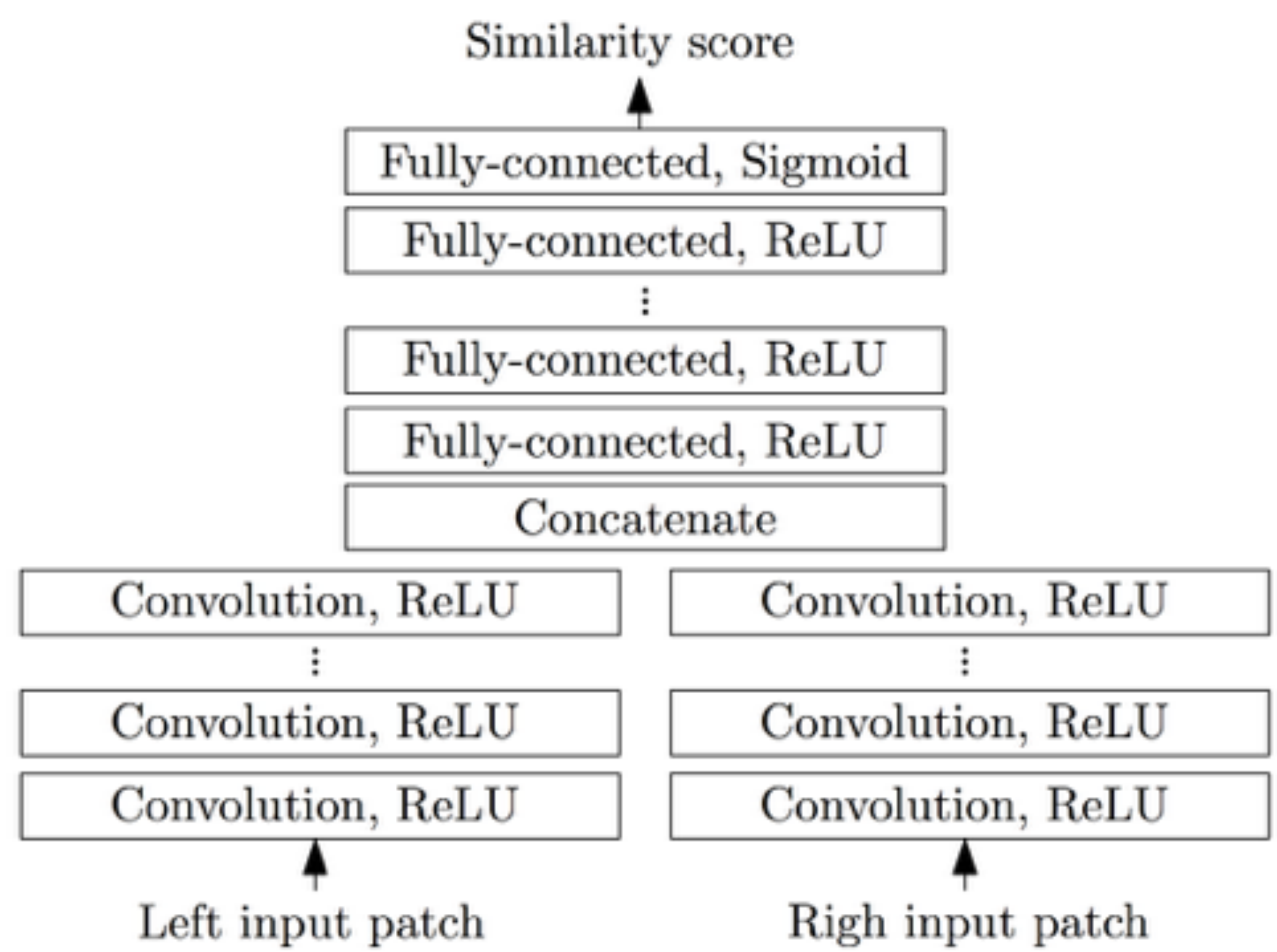
Conv-Nets

- Input: two image patches
  - Equivalent
- Output: matching cost

- What architecture would you use?

Network I

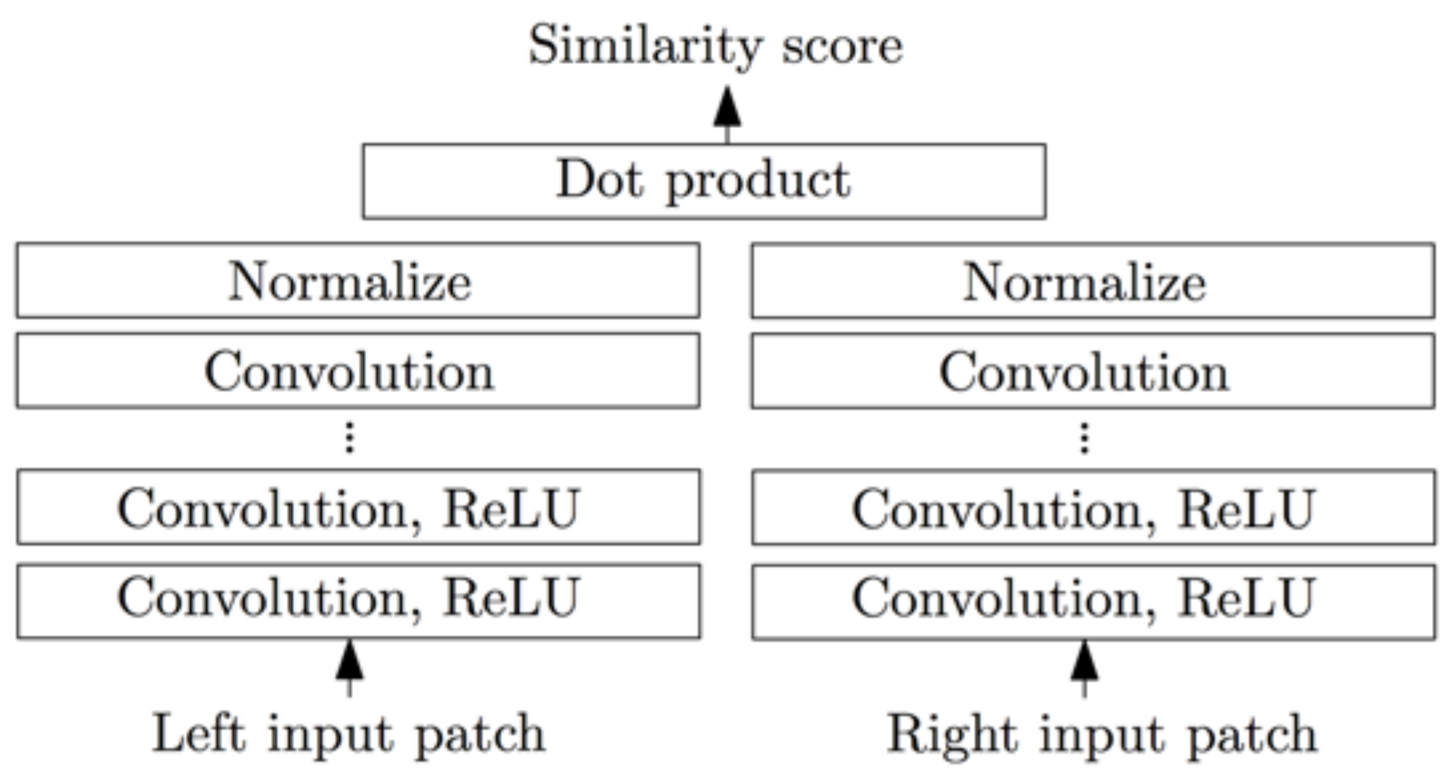
- Two stages:
  - Siamese network
  - Fully connected
- Small patch size
- “Big” network(~600K)
- Binary prediction



Source: Zbontar & LeCun

Network II

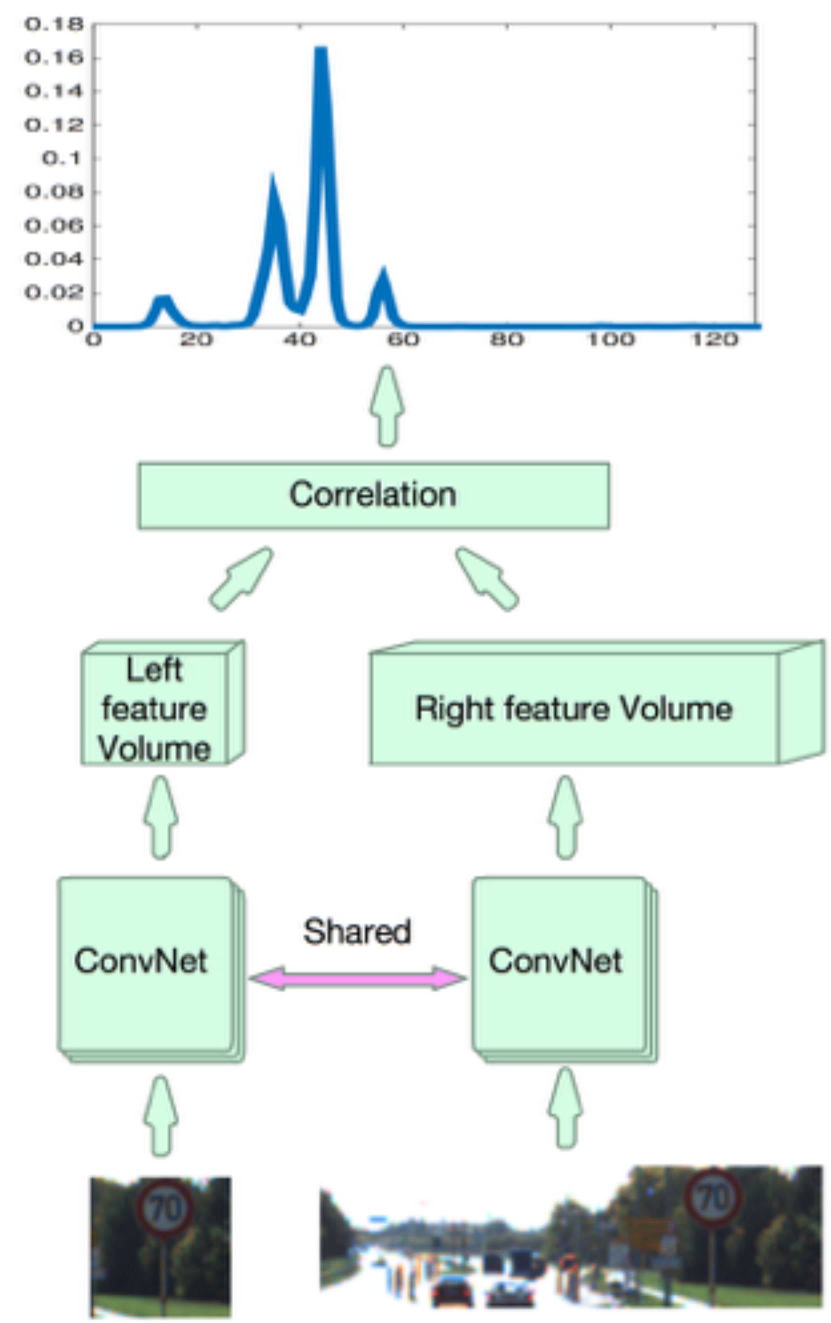
- Dot-product
- Small network
- Hinge loss



Source: Zbontar & LeCun

Network III

- Full content
- Dot-product
- Larger patch
- Log loss



Dataset

Middlebury

KITTI

- Laboratory
- Lambertian
- Rich in texture
- Medium-size label set
- Largely fronto-parallel

- Moving vehicle
- Specularities
- Sensor saturation
- Large label set
- Strong slants

Source: R. Urtasun

Training

- Preprocessing, data-augmentation
- Siamese network: gradient aggregated
- SGD; Batch Normalization

Test

- Image size: W, H; Disparity range: D
    - W * H * D: $1200 \times 370 \times 256 = 1.14 \times 10^8!$
- Computation
    - Feature shared
- Memory
    - One disparity at a time

Smoothing

- Cost-aggregation
  - Averaging neighboring locations
  - Fancy “neighborhood”
- CRF
  - What energy would you use?

CRF

- Minimize energy:

$$E(y) = \sum_{i=1}^{N} E_i(y_i) + \sum_{(i,j)\in E} E_{i,j}(y_i, y_j)$$

$E_i(y_i)$: energy of unary potential; $E_{i,j}(y_i, y_j)$: energy on edge

SGM

- Potential:

$$E_{i,j}(y_i, y_j) = \begin{cases} 0 & \text{if } y_i = y_j \\ c_1 & \text{if } |y_i - y_j| = 1 \\ c_2 & \text{otherwise} \end{cases}$$

- Global optimum: NP-hard

- One direction with dynamic programming:

$$O(W \cdot H \cdot D)$$

- Averaging over multiple directions

Slanted plane

- Continuity/smoothness within a [slanted] plane

$$d(\mathbf{p}, \theta_i) = A_i p_x + B_i p_y + C_i,\ \ \theta_i = (A_i, B_i, C_i)$$

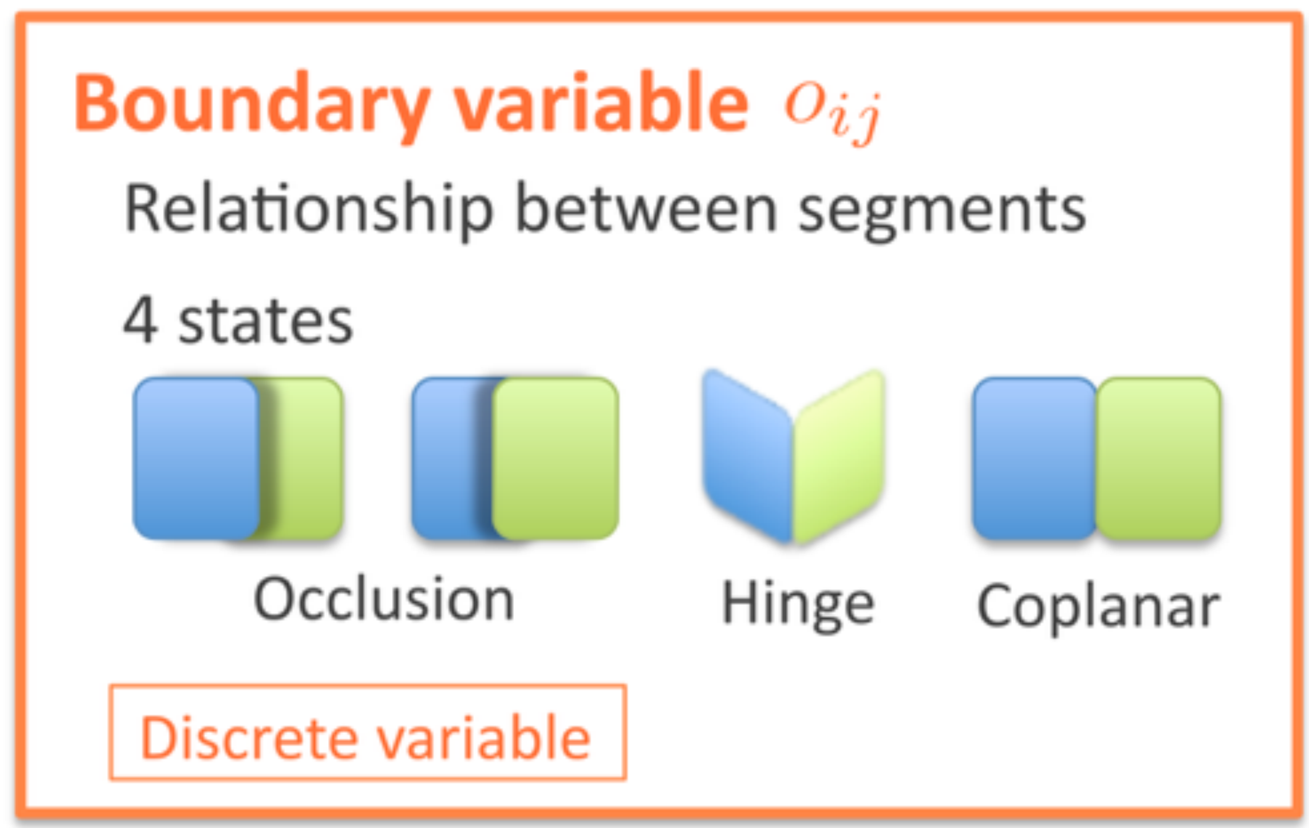
- What energy term? (Pixel, Segment, Plane)
    - pixel & segment: color, location
    - pixel & plane: disparity
    - segment: boundary length

Slanted plane cont.

- Segment & plane
  - complexity(prior):

    co-planar > hinge > occlusion

  - boundary-plane consistency



Source: R. Urtasun

Refinement

- Border fixing(CNN)
- Left-right consistency
- Further smooth
- Outlier detector

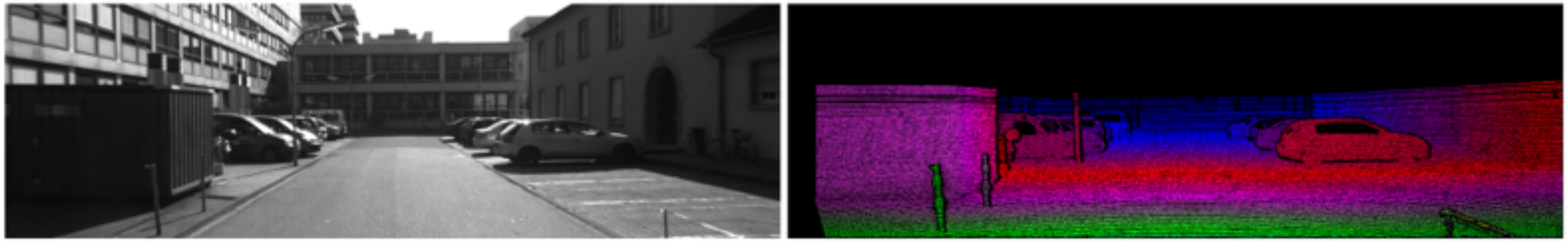
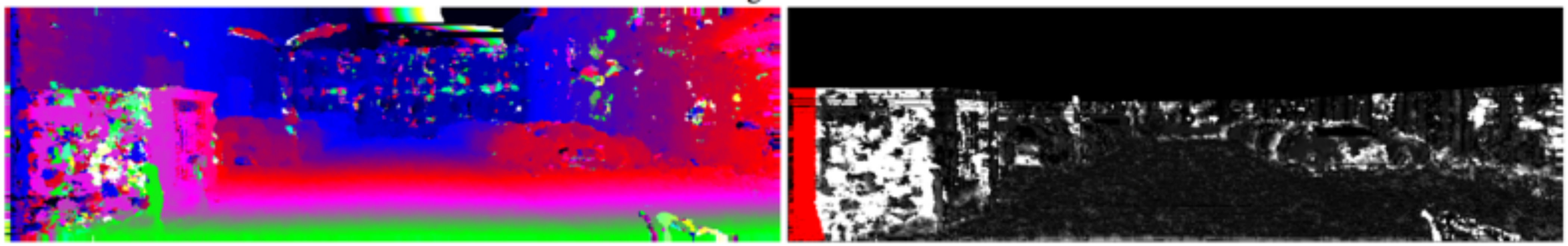
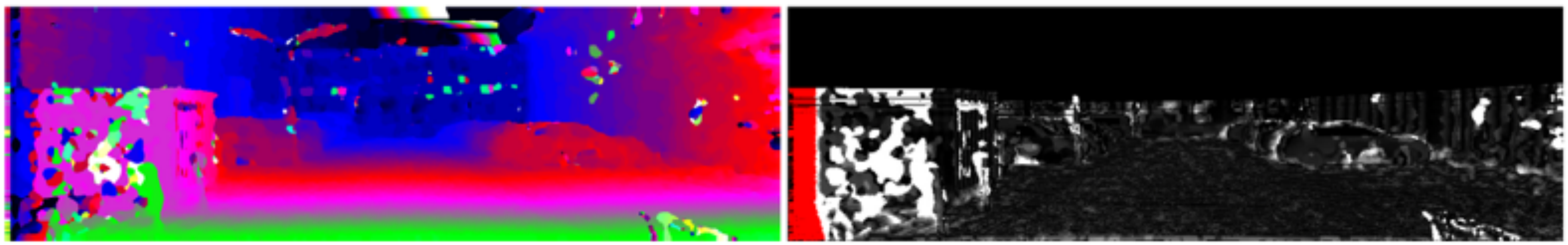


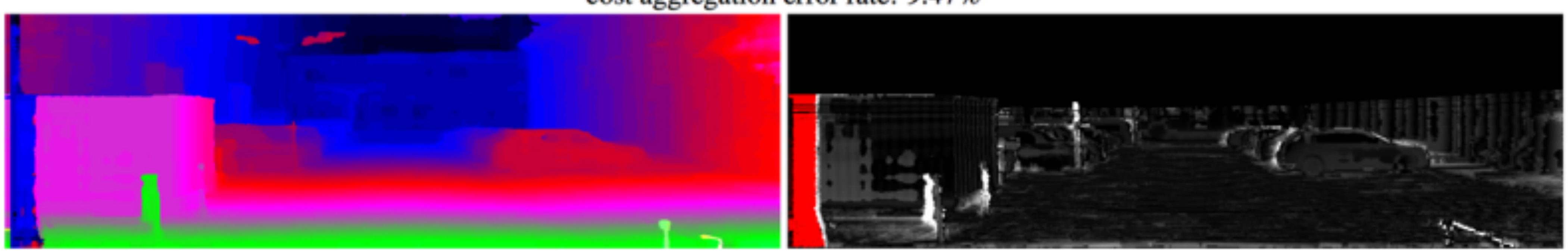
image id: 170



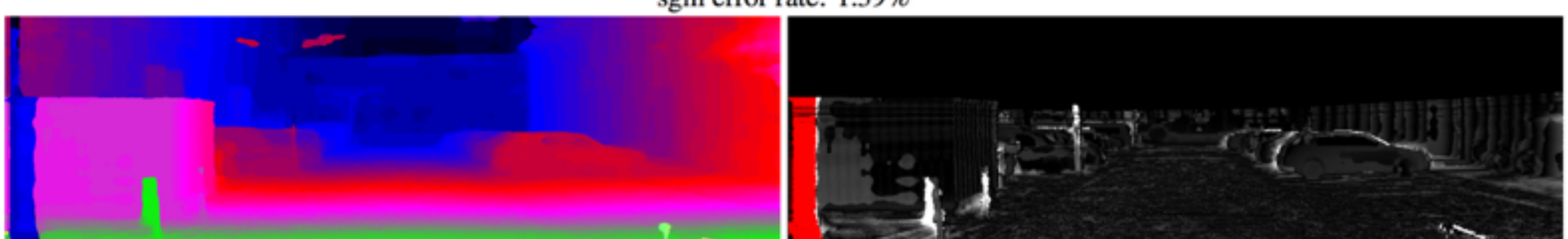
cnn error rate: 13.48%



cost aggregation error rate: 9.47%



sgm error rate: 1.39%



final error rate: 1.15%

What else?

- Better CRF & inference
- End-to-End training
- Joint with segmentation

Thank You

Q&A