Depth and Surface Normal Estimation from a Single Image

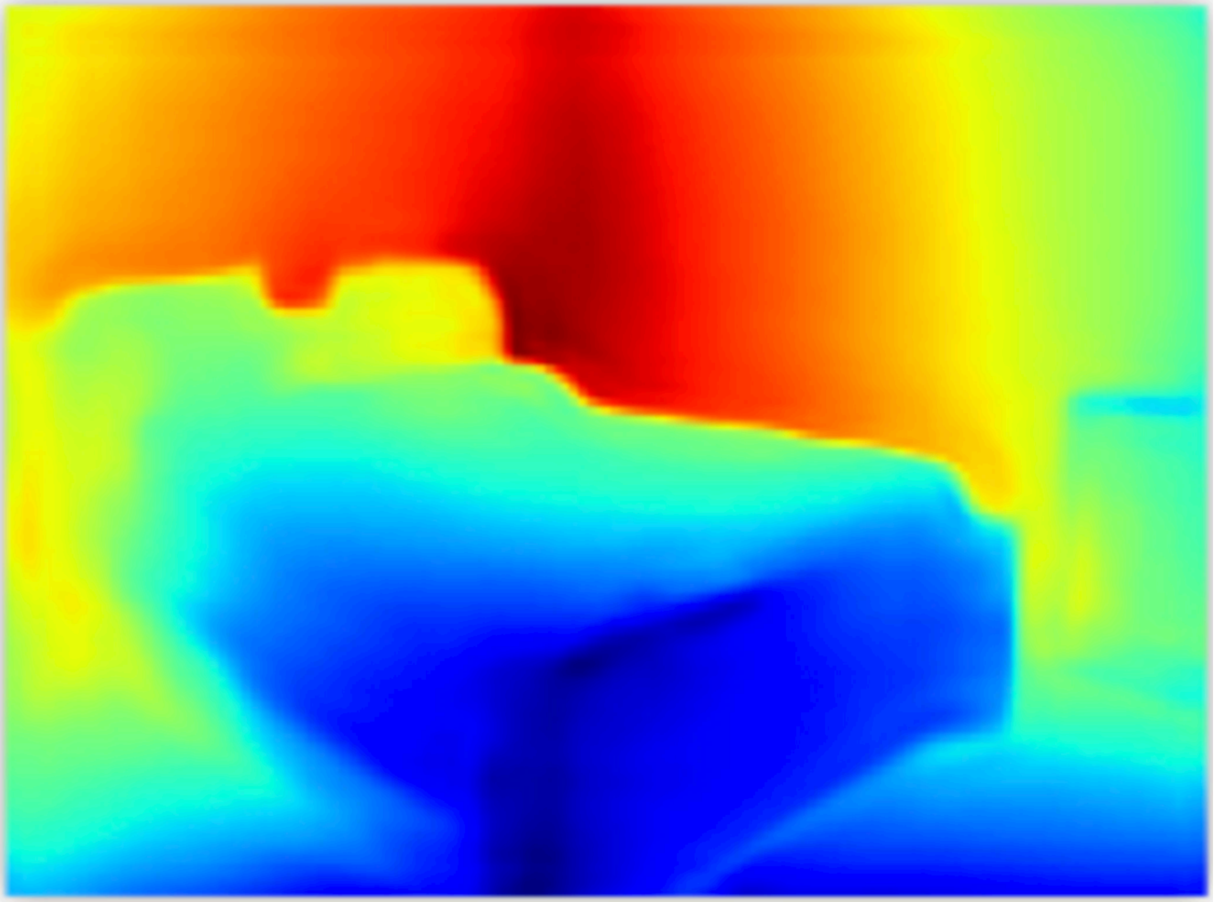
Mian Wei
University of Toronto

What is the problem?

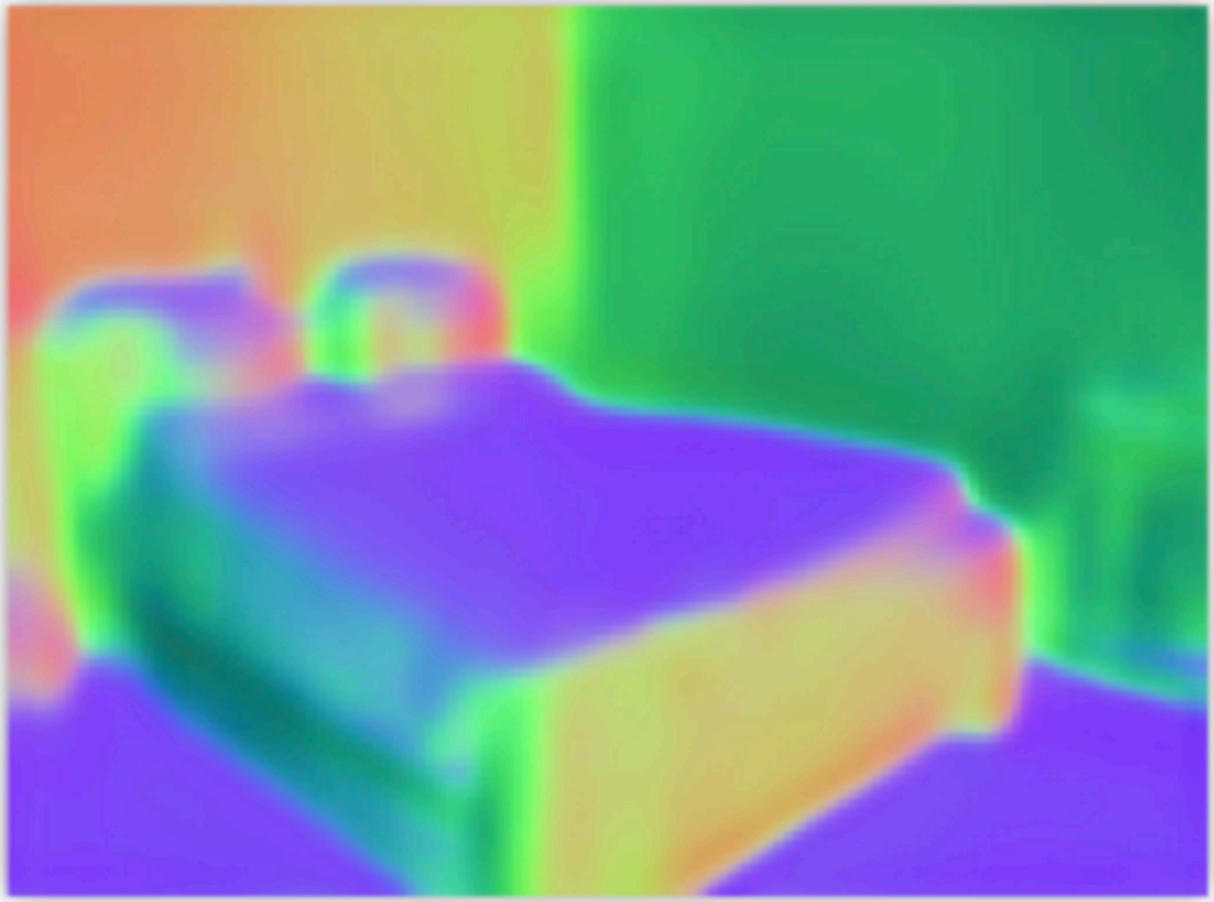
Given one image

N. Silberman, D. Hoiem, P. Kohli, and R. Fergus, "Indoor segmentation and support inference from RGBD images," in *Proc. Eur. Conf. Comput. Vision*, 2012, pp. 746–760.

Estimate the following:



Eigen, D. and Fergus, R. Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture. ICCV 2015

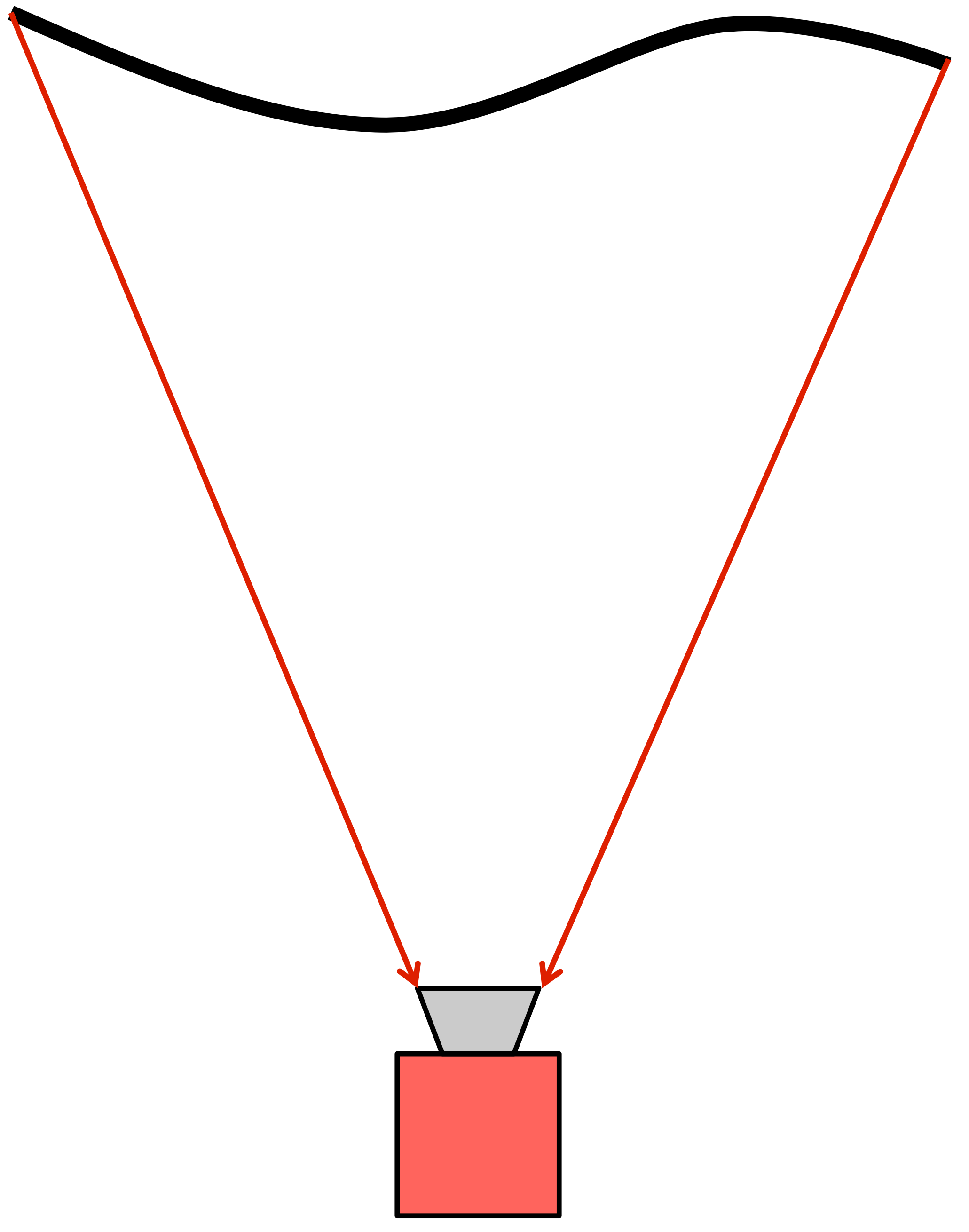


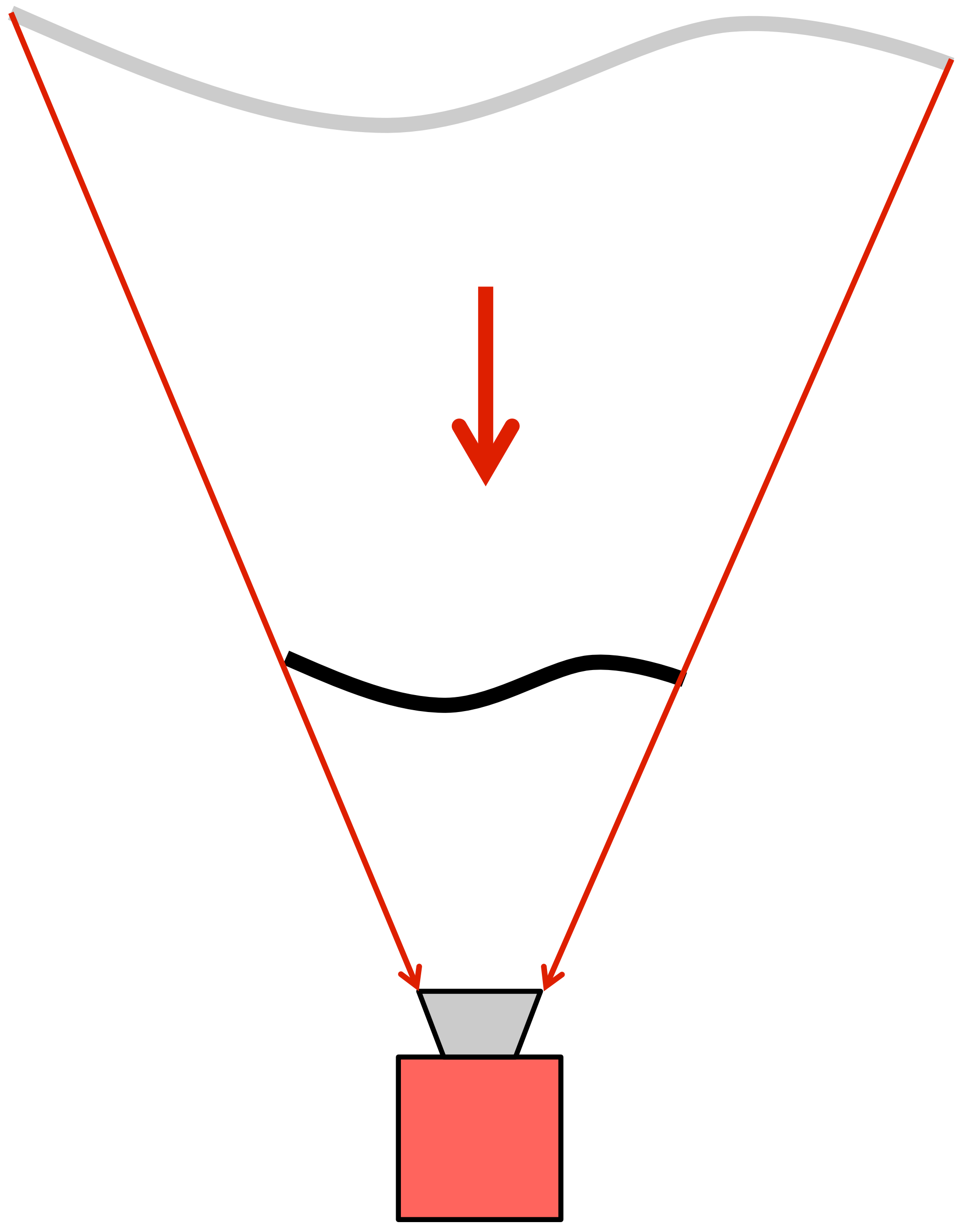
Eigen, D. and Fergus, R. Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture. ICCV 2015

Why is this hard?

Multiple ambiguities

Scale ambiguity



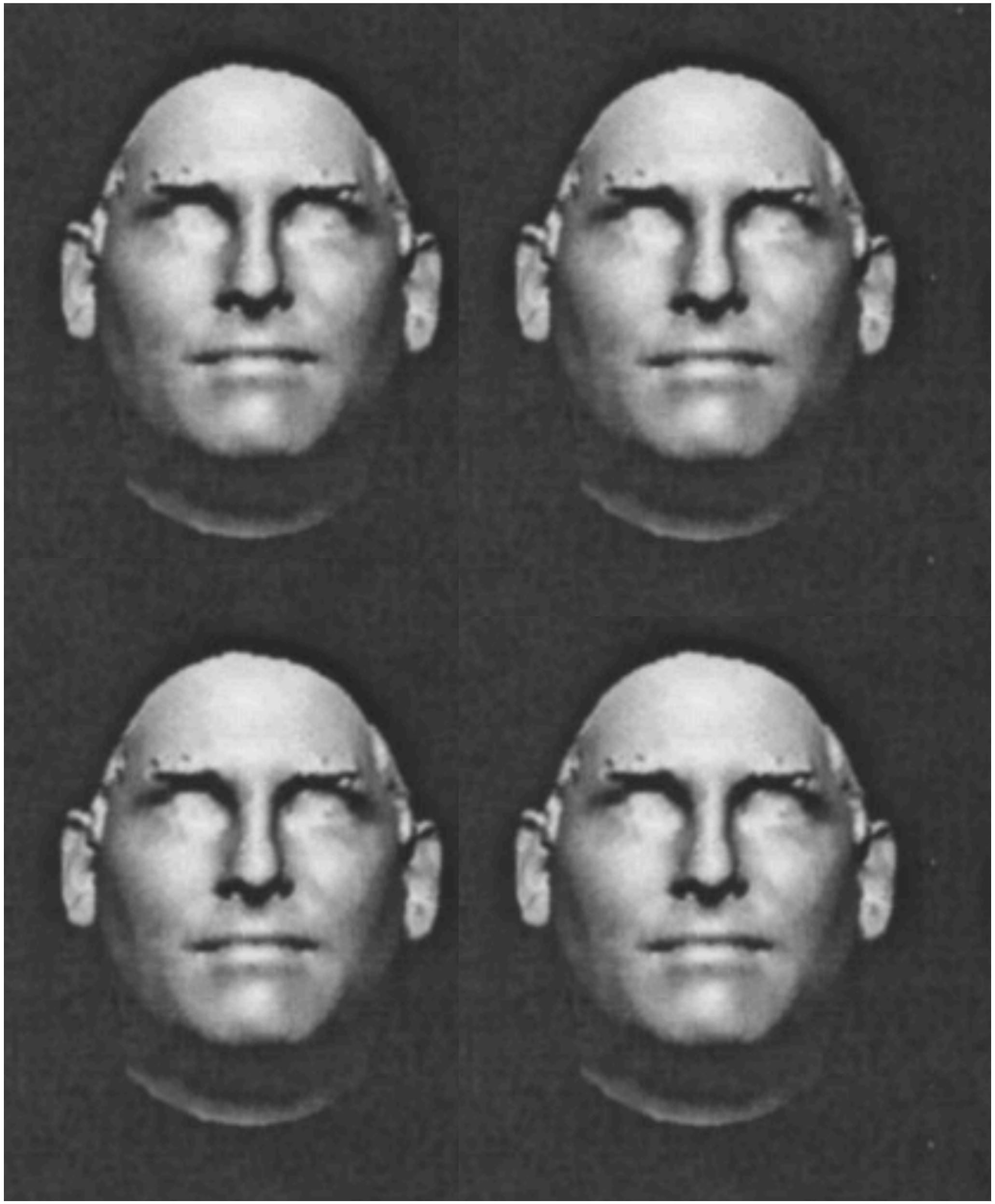


Bas-relief ambiguity

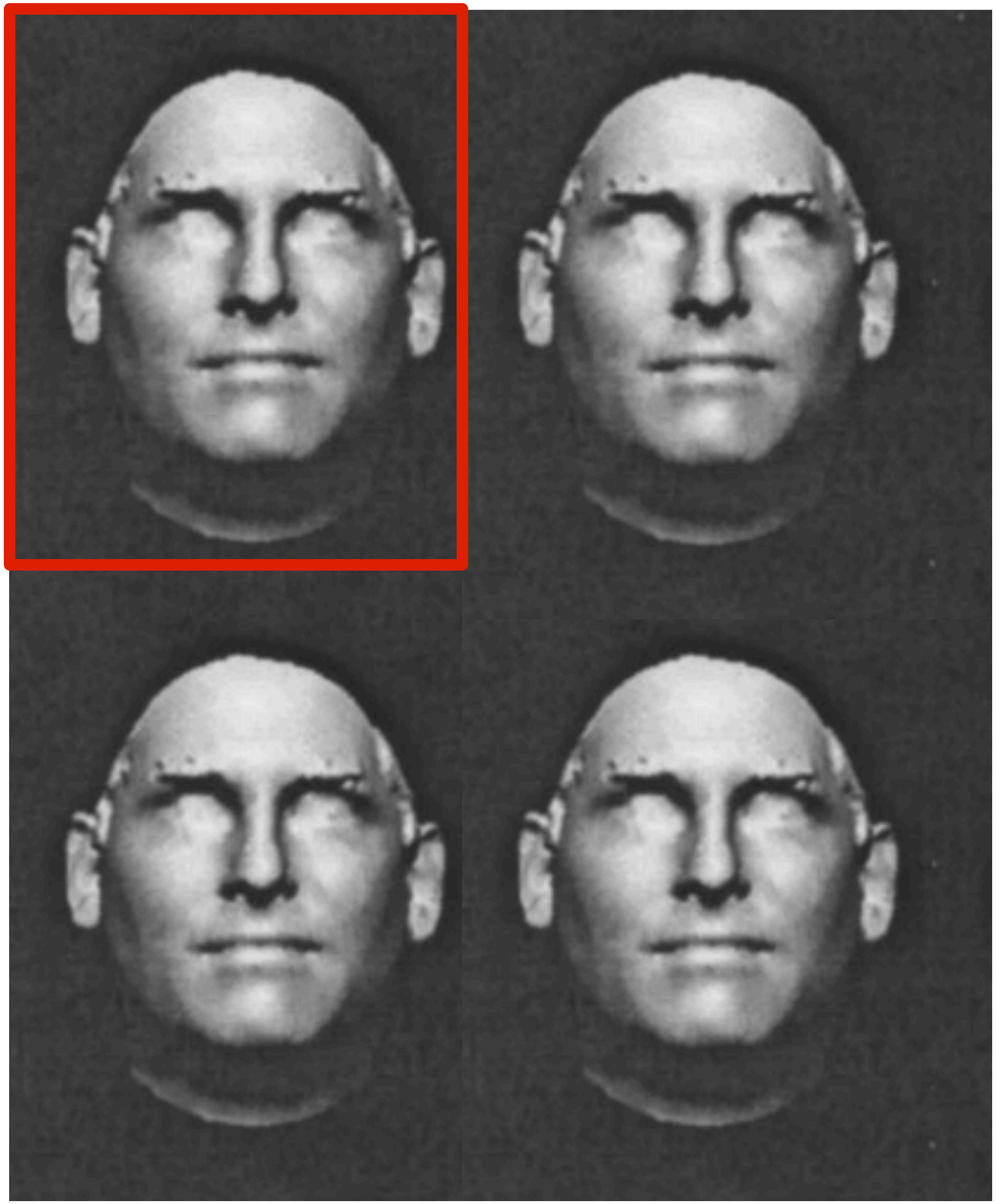
P. Belhumeur, D. Kriegman, and A. Yuille, "The Bas-Relief Ambiguity," Proc. IEEE Conf. Computer Vision and Pattern Recognition, pp. 1040-1046, 1997.

Let’s play a game

Spot the Difference

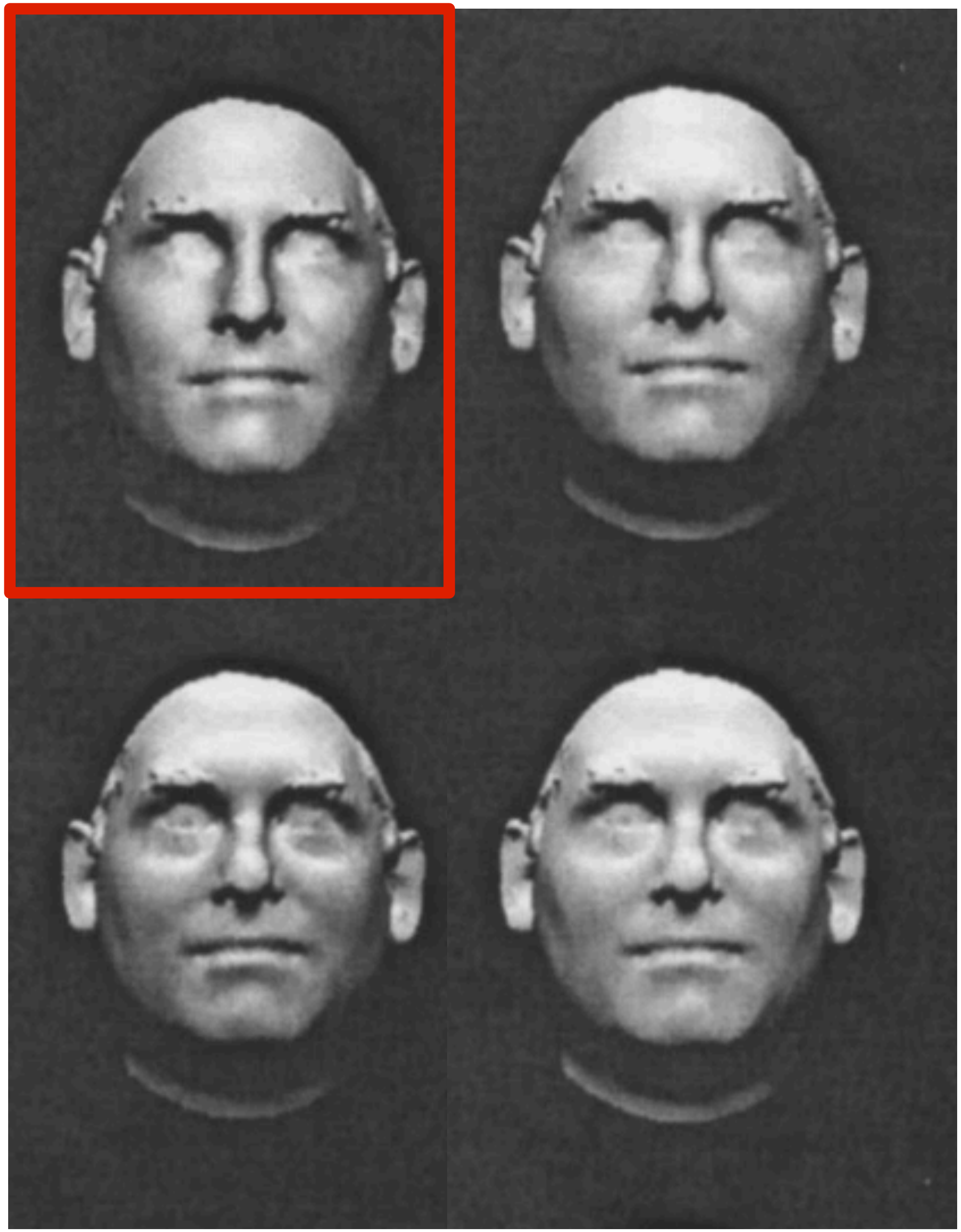


P. Belhumeur, D. Kriegman, and A. Yuille, "The Bas-Relief Ambiguity," Proc. IEEE Conf. Computer Vision and Pattern Recognition, pp. 1040-1046, 1997.

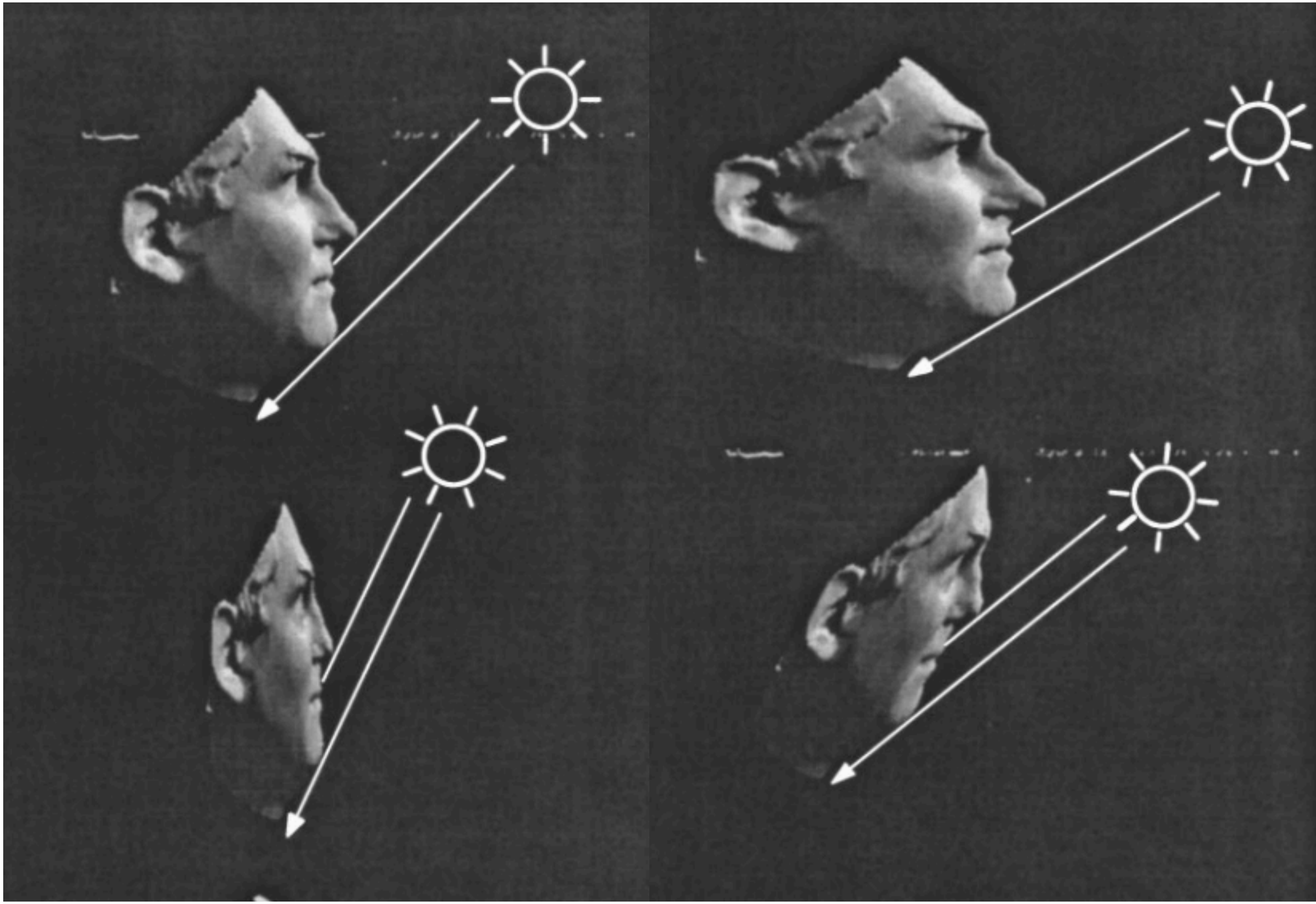


P. Belhumeur, D. Kriegman, and A. Yuille, "The Bas-Relief Ambiguity," Proc. IEEE Conf. Computer Vision and Pattern Recognition, pp. 1040-1046, 1997.

All the same



P. Belhumeur, D. Kriegman, and A. Yuille, "The Bas-Relief Ambiguity," Proc. IEEE Conf. Computer Vision and Pattern Recognition, pp. 1040-1046, 1997.



P. Belhumeur, D. Kriegman, and A. Yuille, "The Bas-Relief Ambiguity," Proc. IEEE Conf. Computer Vision and Pattern Recognition, pp. 1040-1046, 1997.

Family of transformation

Generalized Bas-Relief

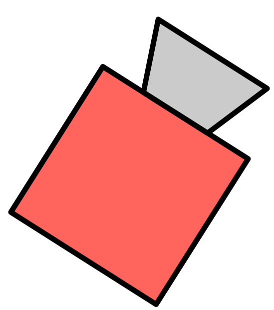
Change shape and illumination

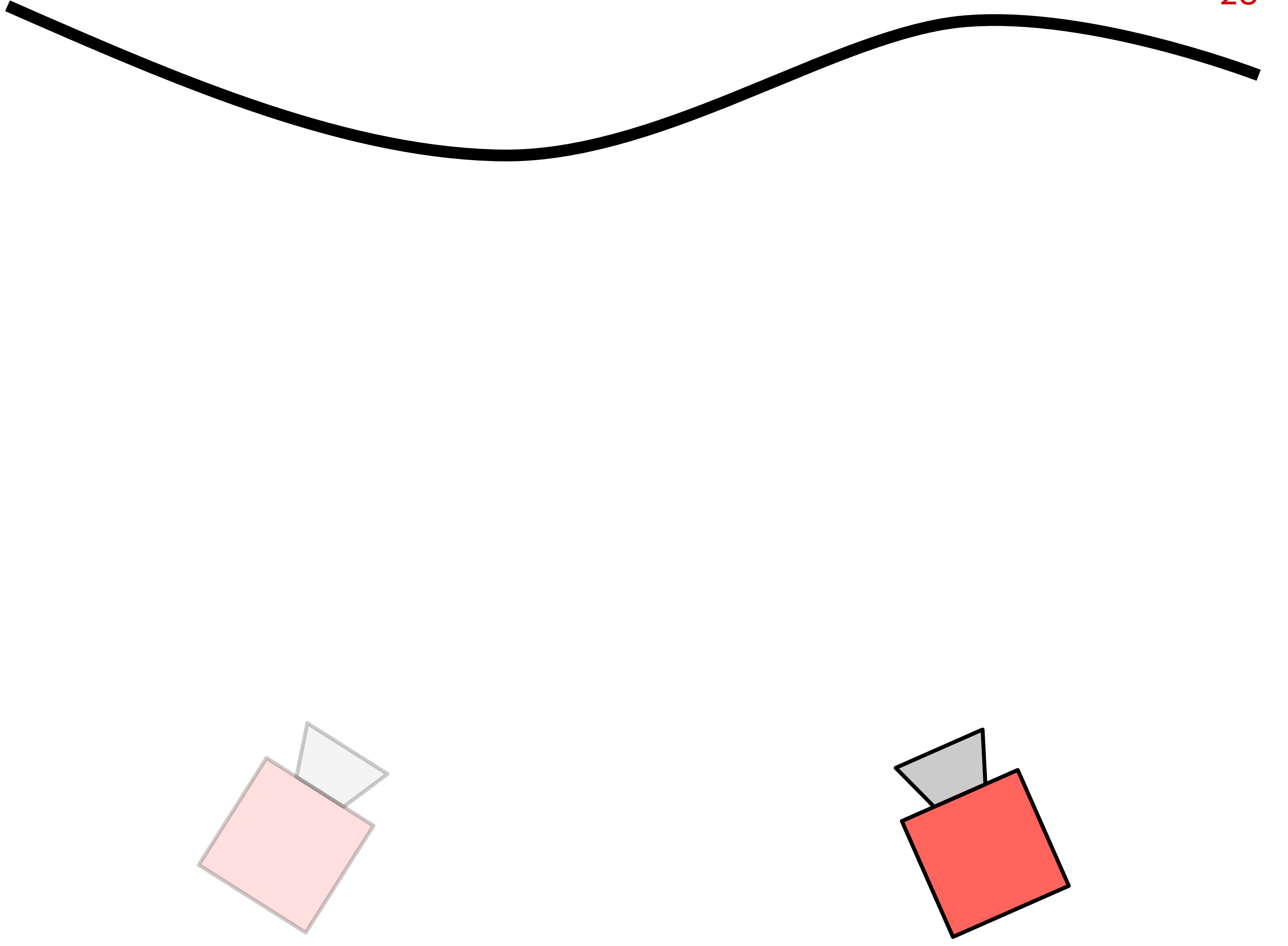
Yield same image

Existing works

Multi-view Stereo

Hartley, R. and Zisserman, A. 2000. *Multiple view geometry in computer vision*, Cambridge University Press: Cambridge, UK.



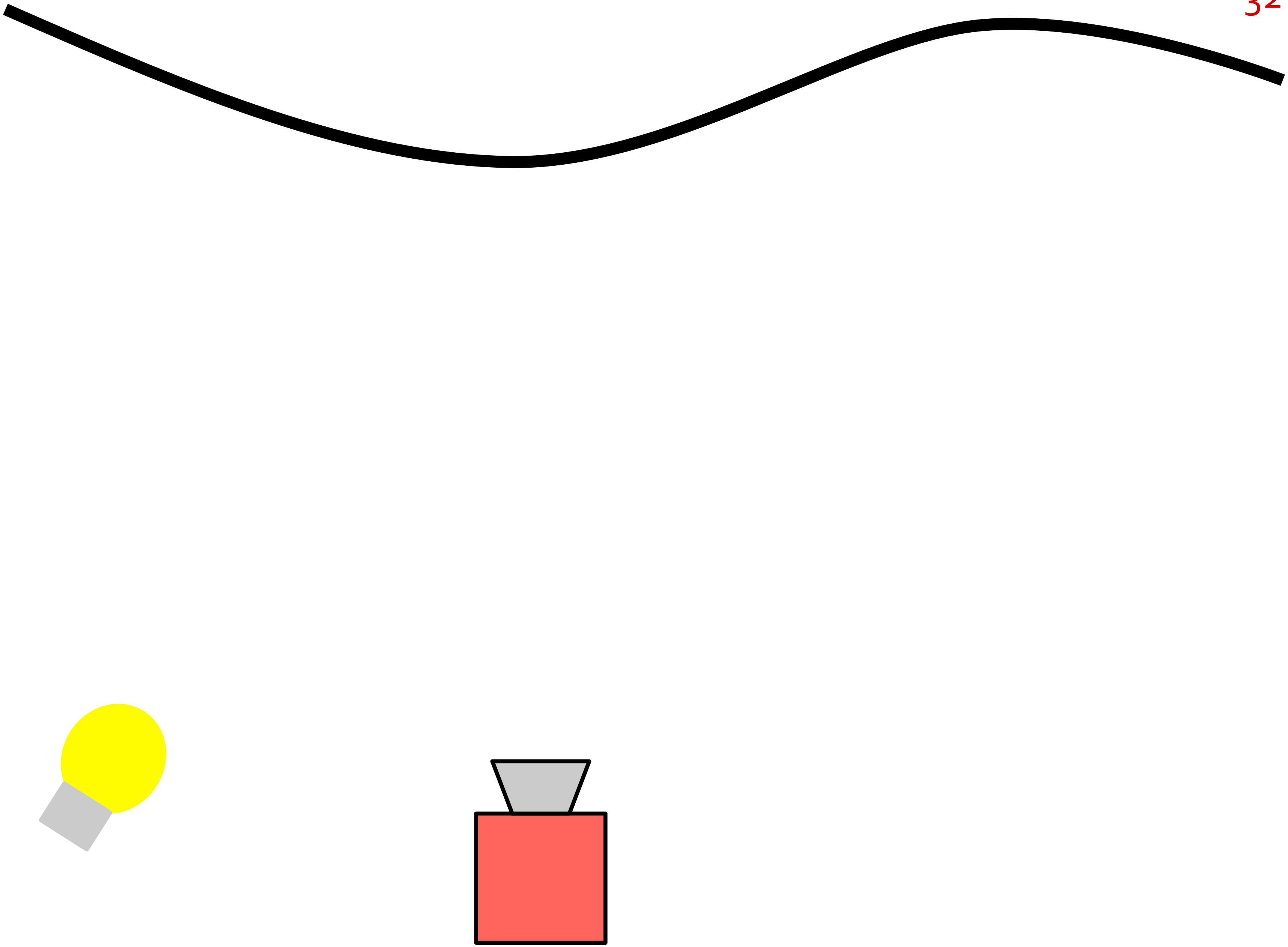


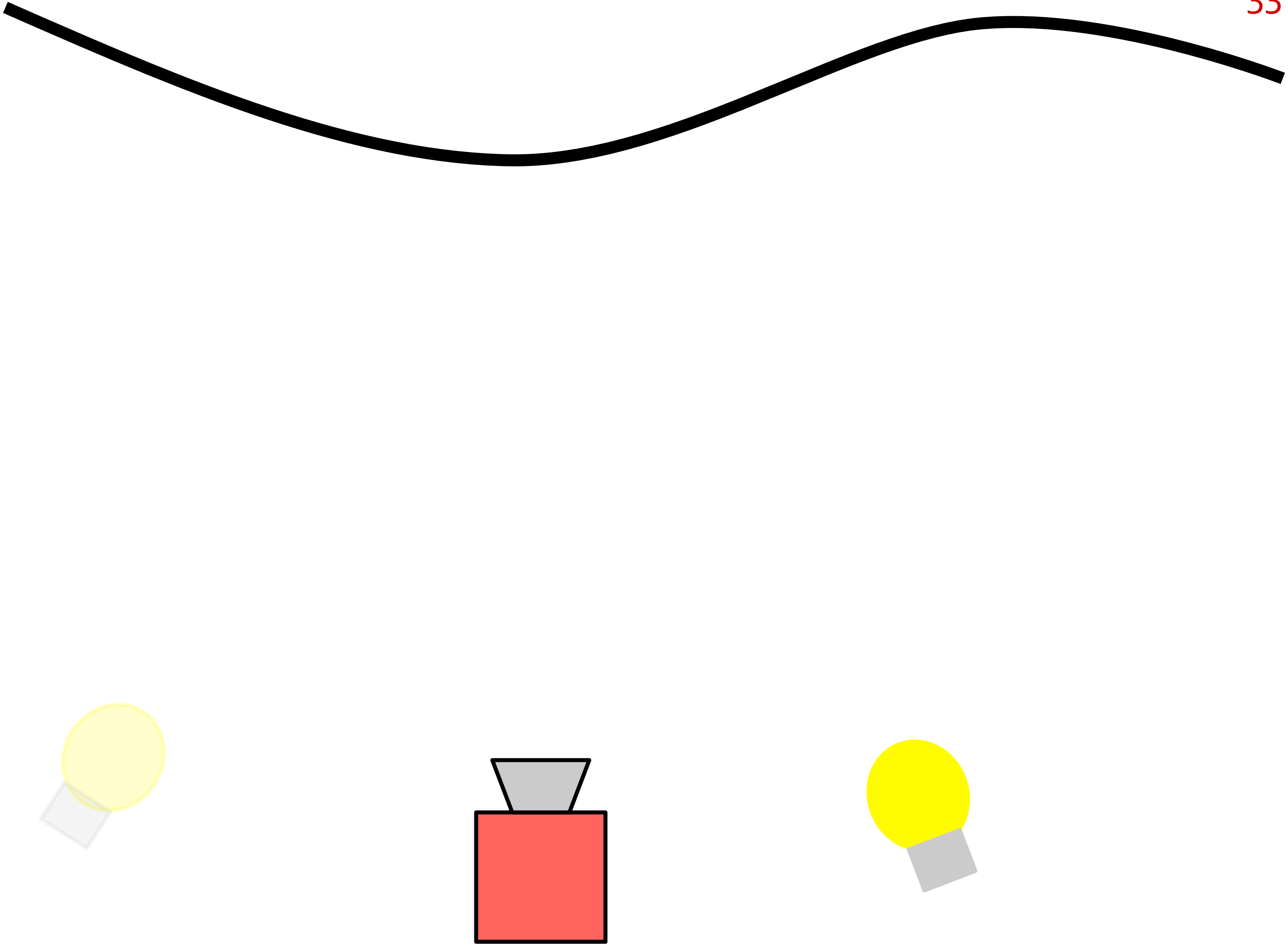
Photometric Stereo

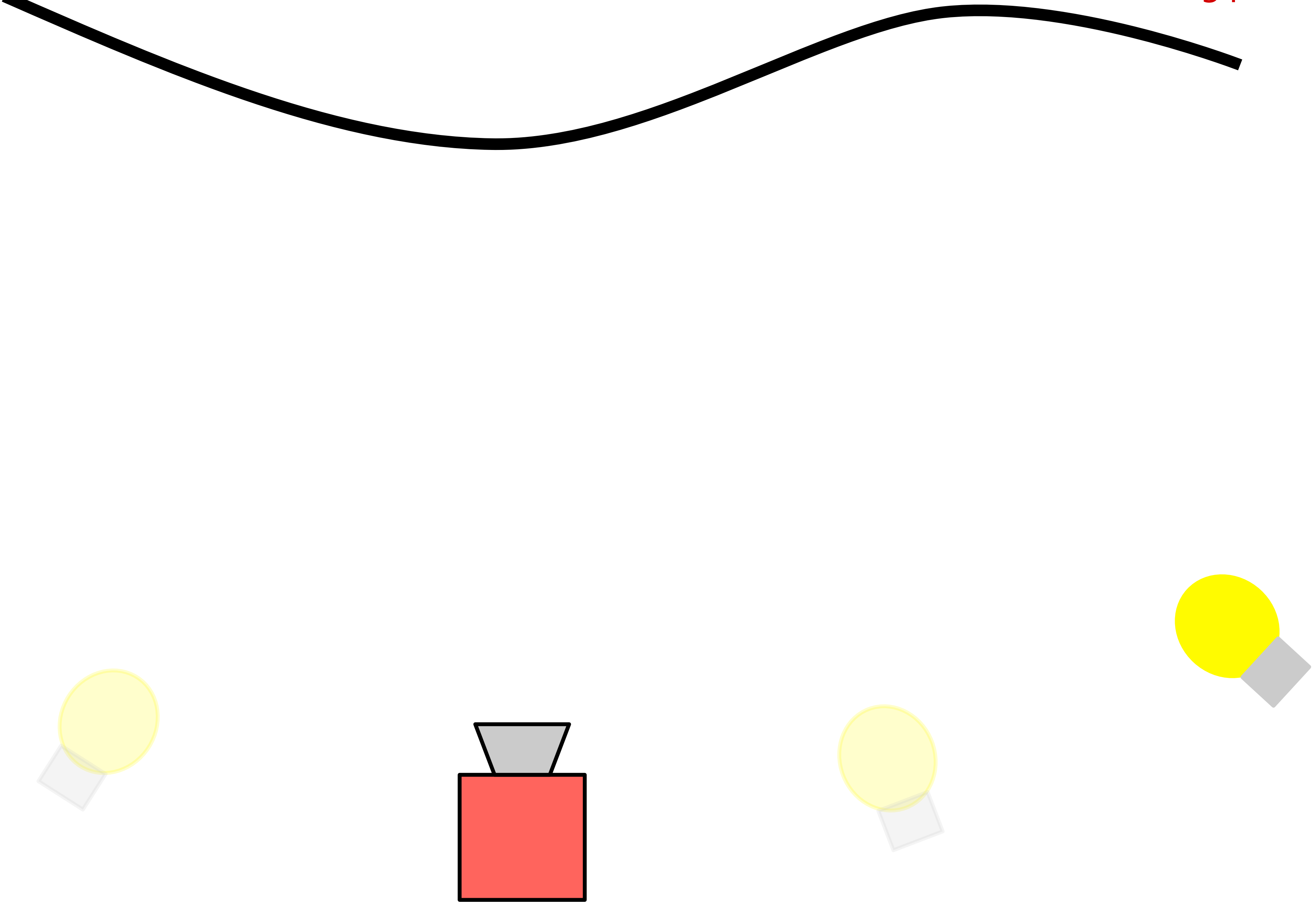
Woodham, R.J. (1980), Photometric method for determining surface orientation from multiple images, Optical Engineering 19 (1) 139-144.

Collimated Light Sources

Light rays parallel

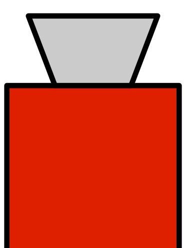


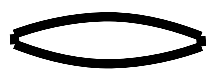


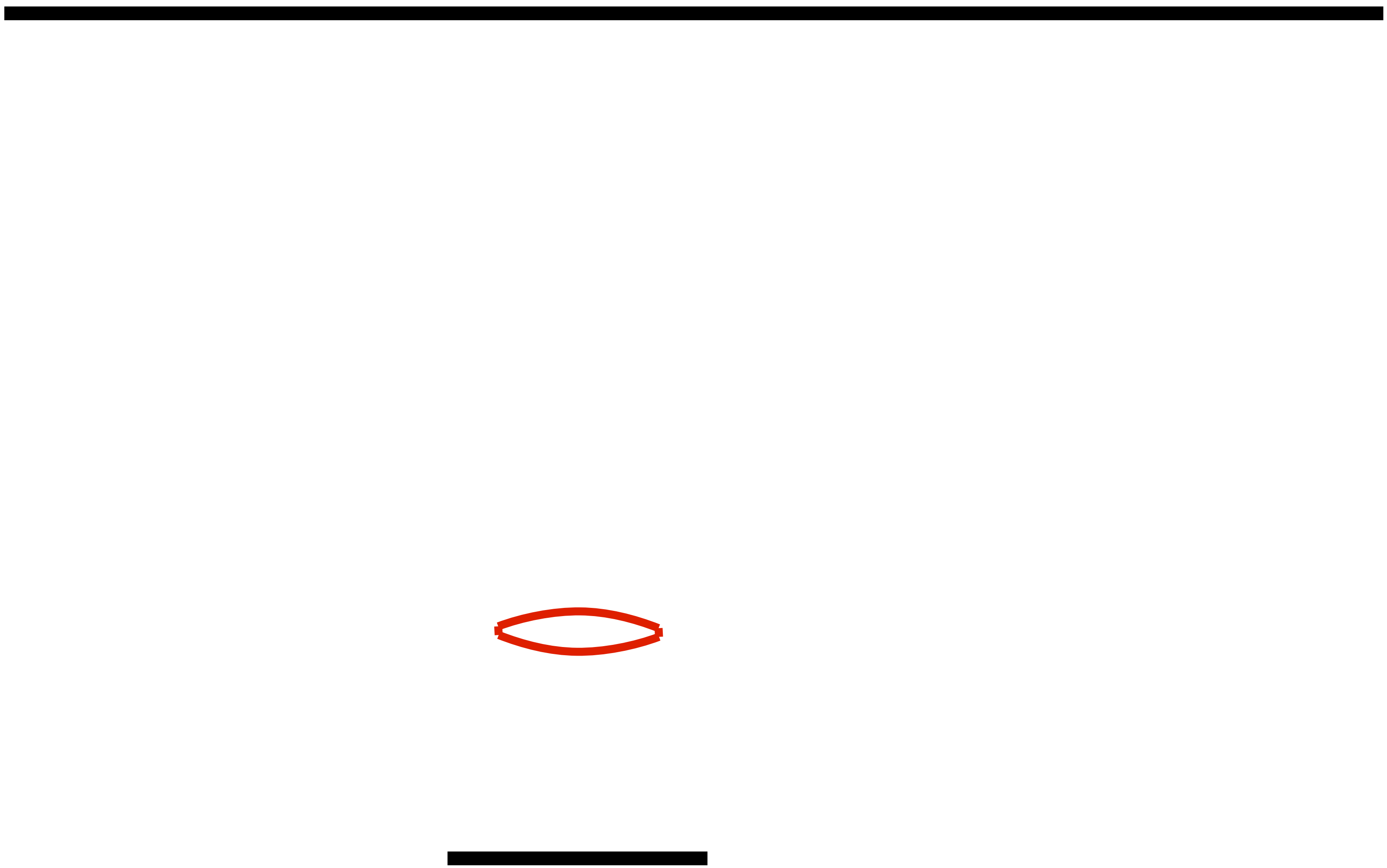


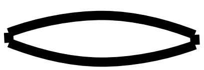
Shape from Focus

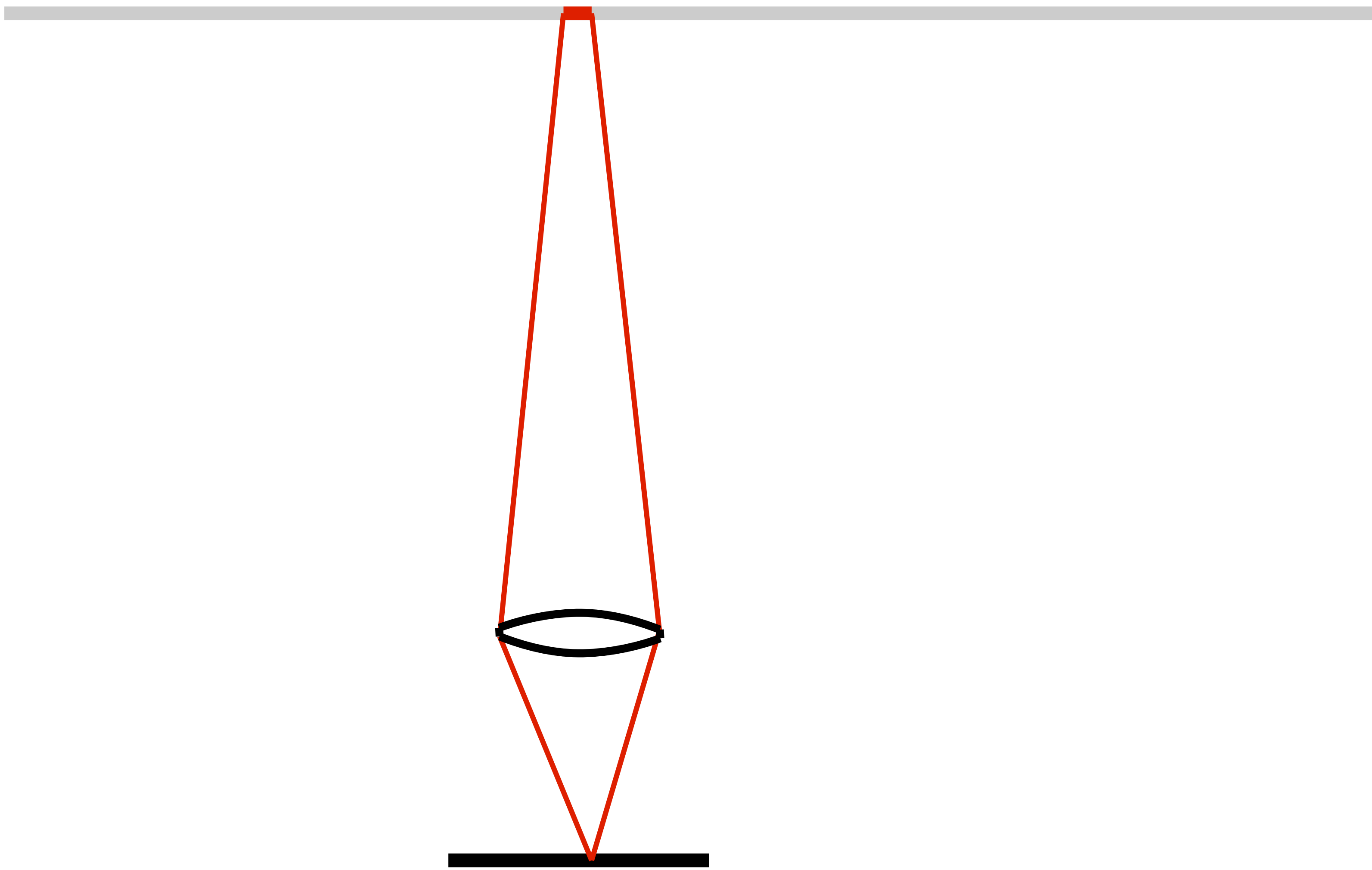
S. Nayar and N. Yasuo, "Shape From Focus," IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 16, no. 8, pp. 824-831, 1994.

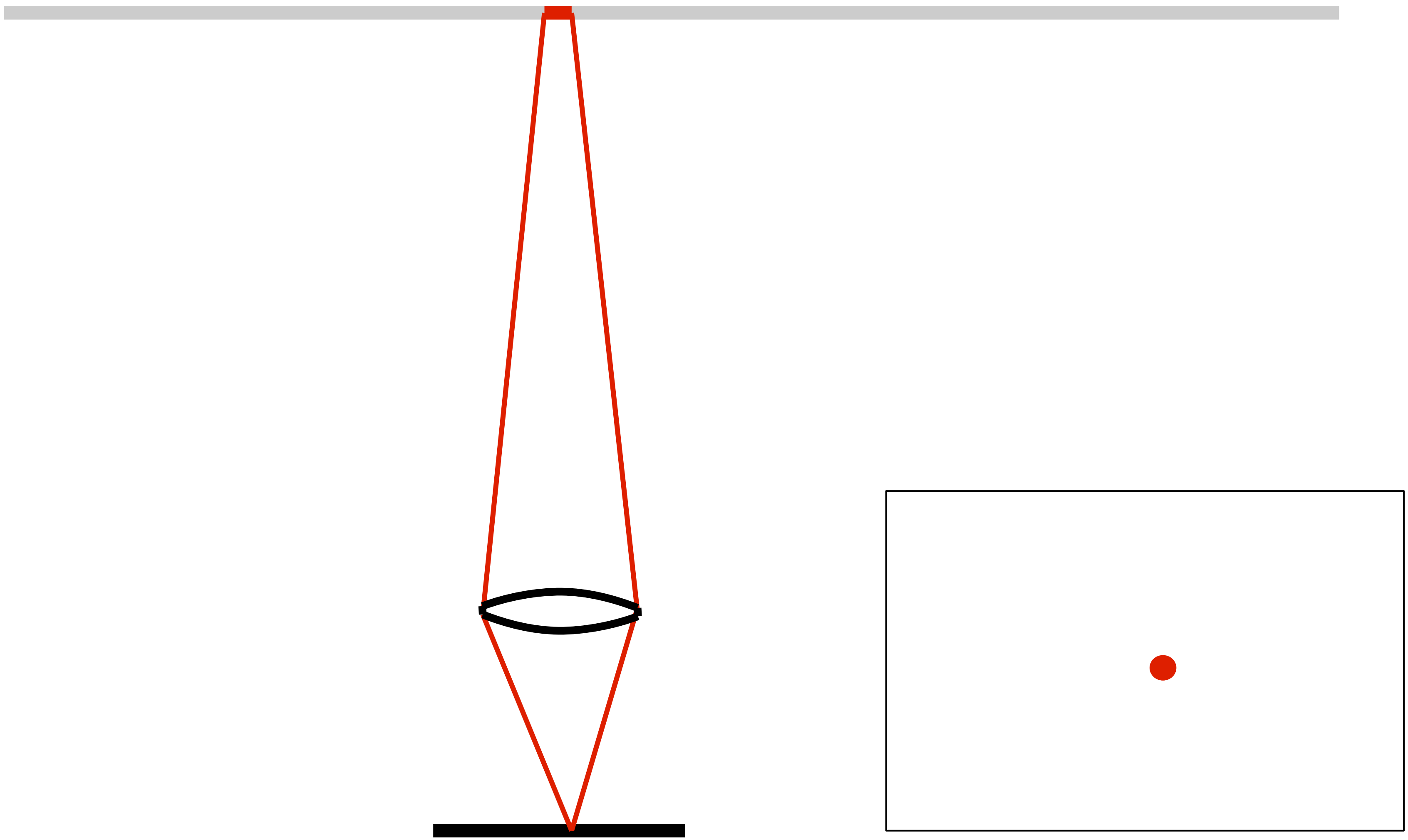


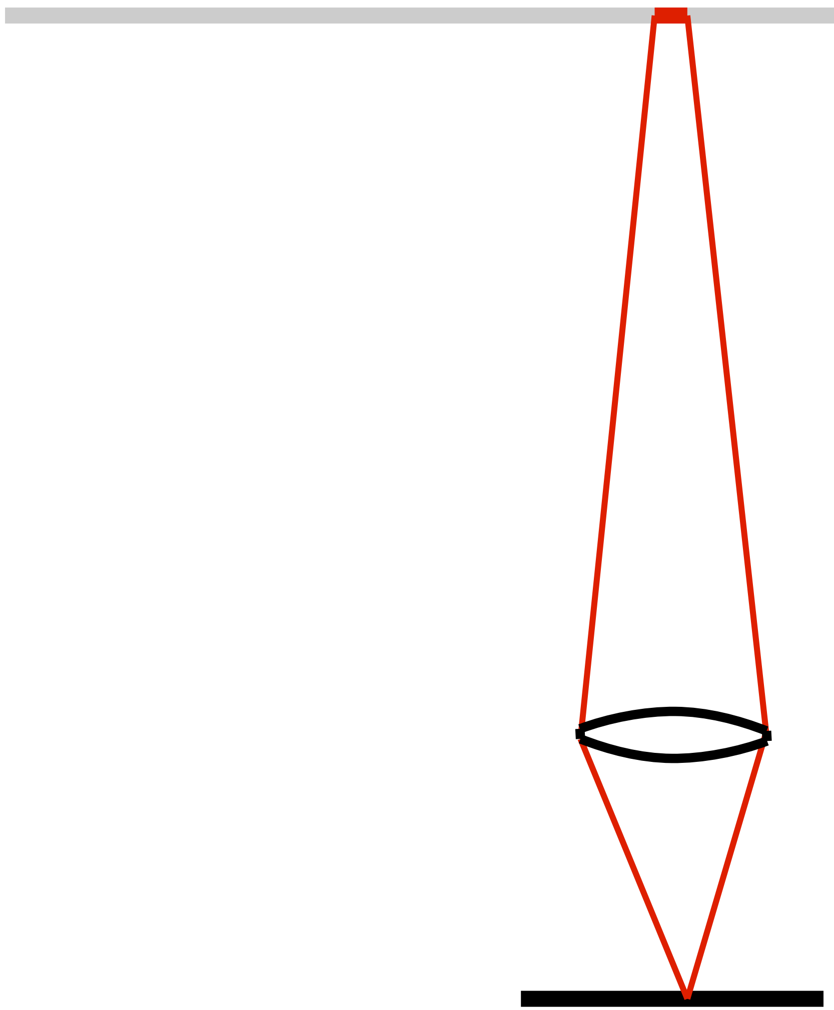


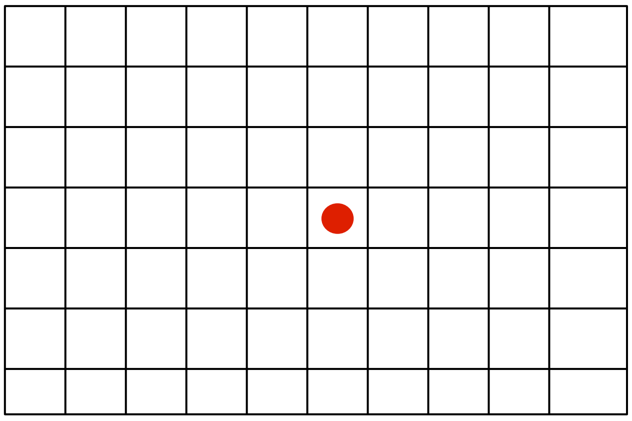


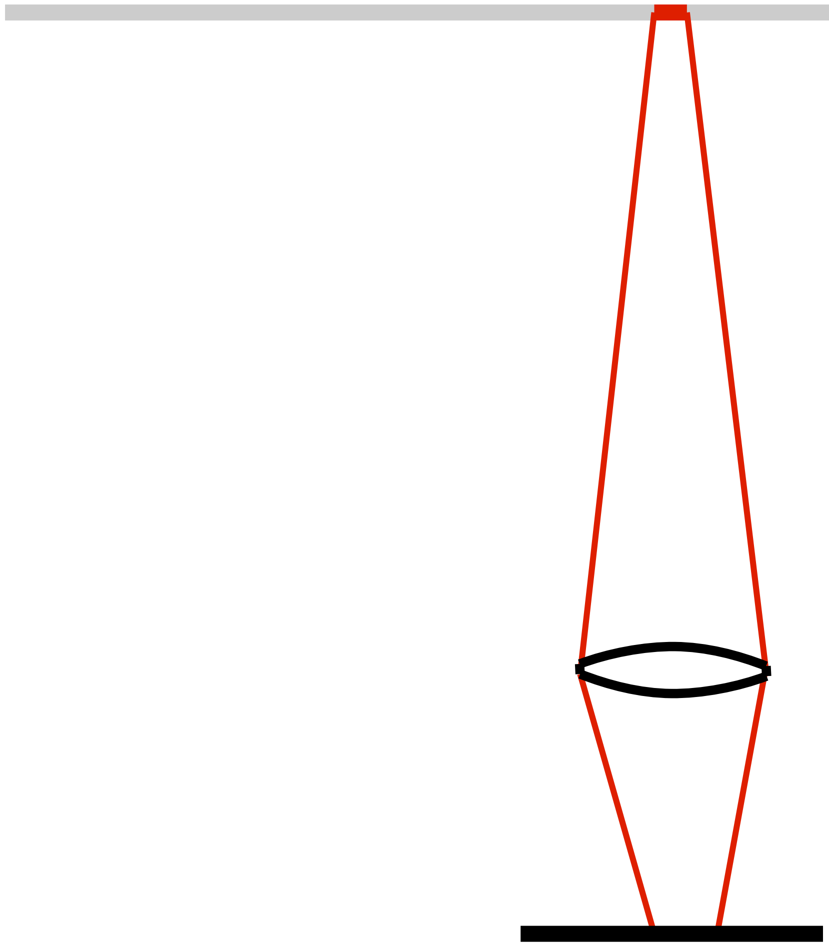


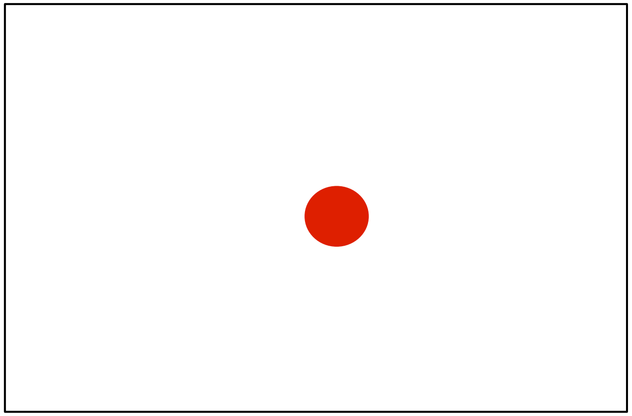


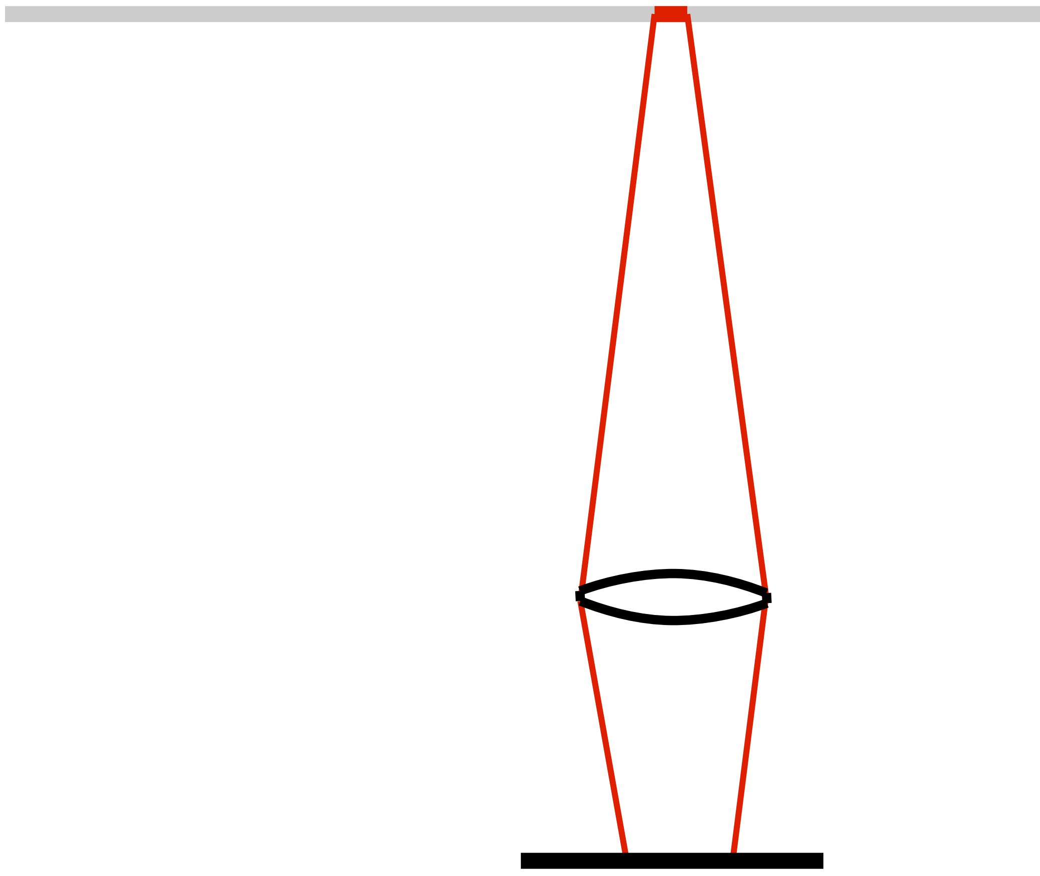


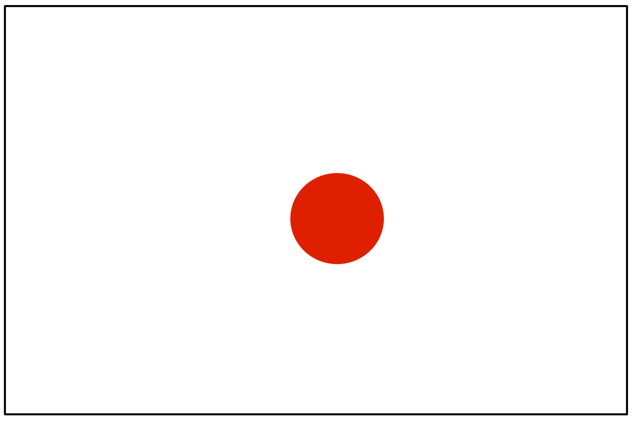


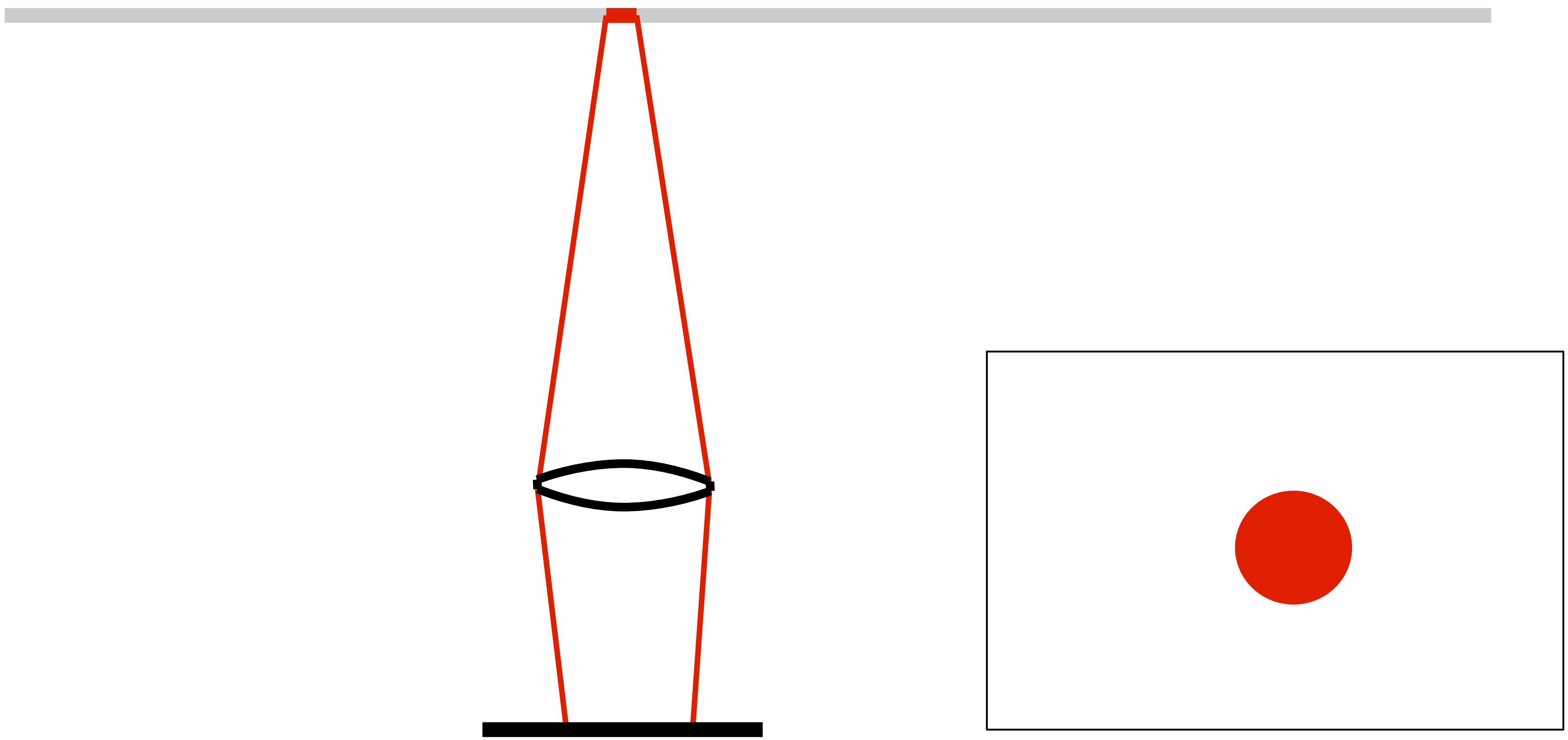


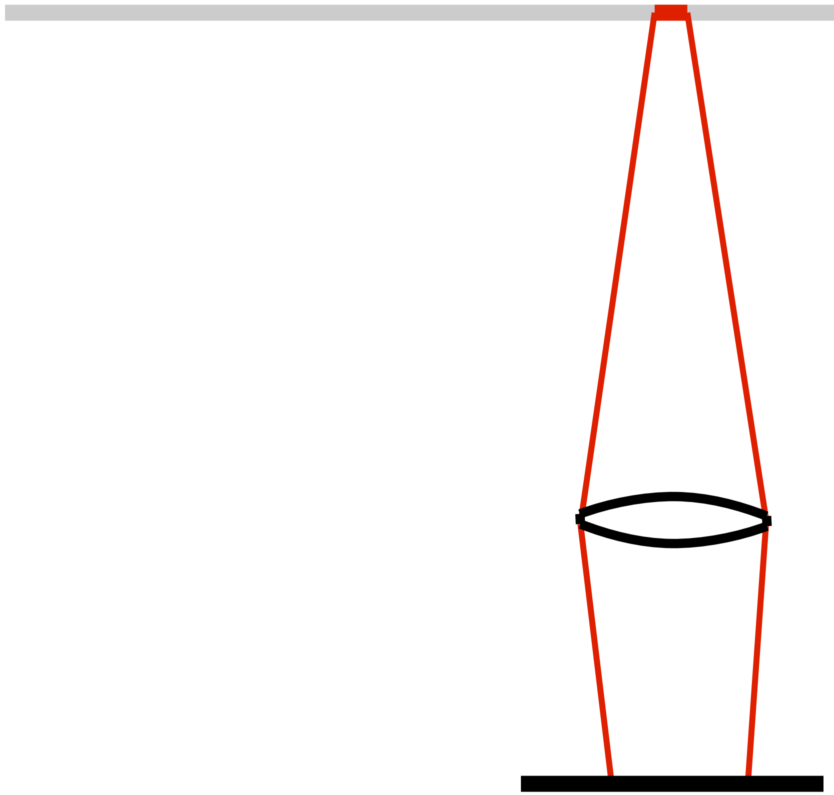


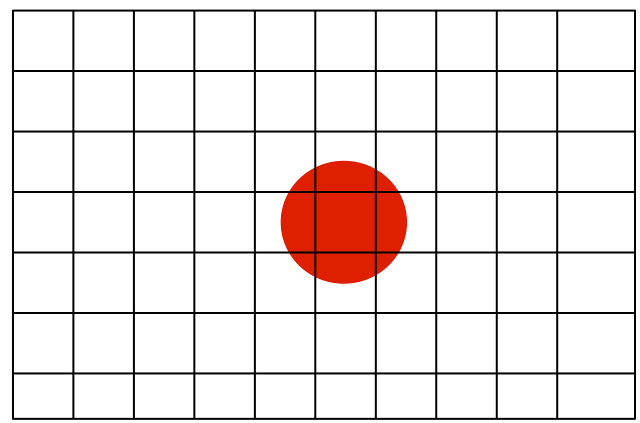






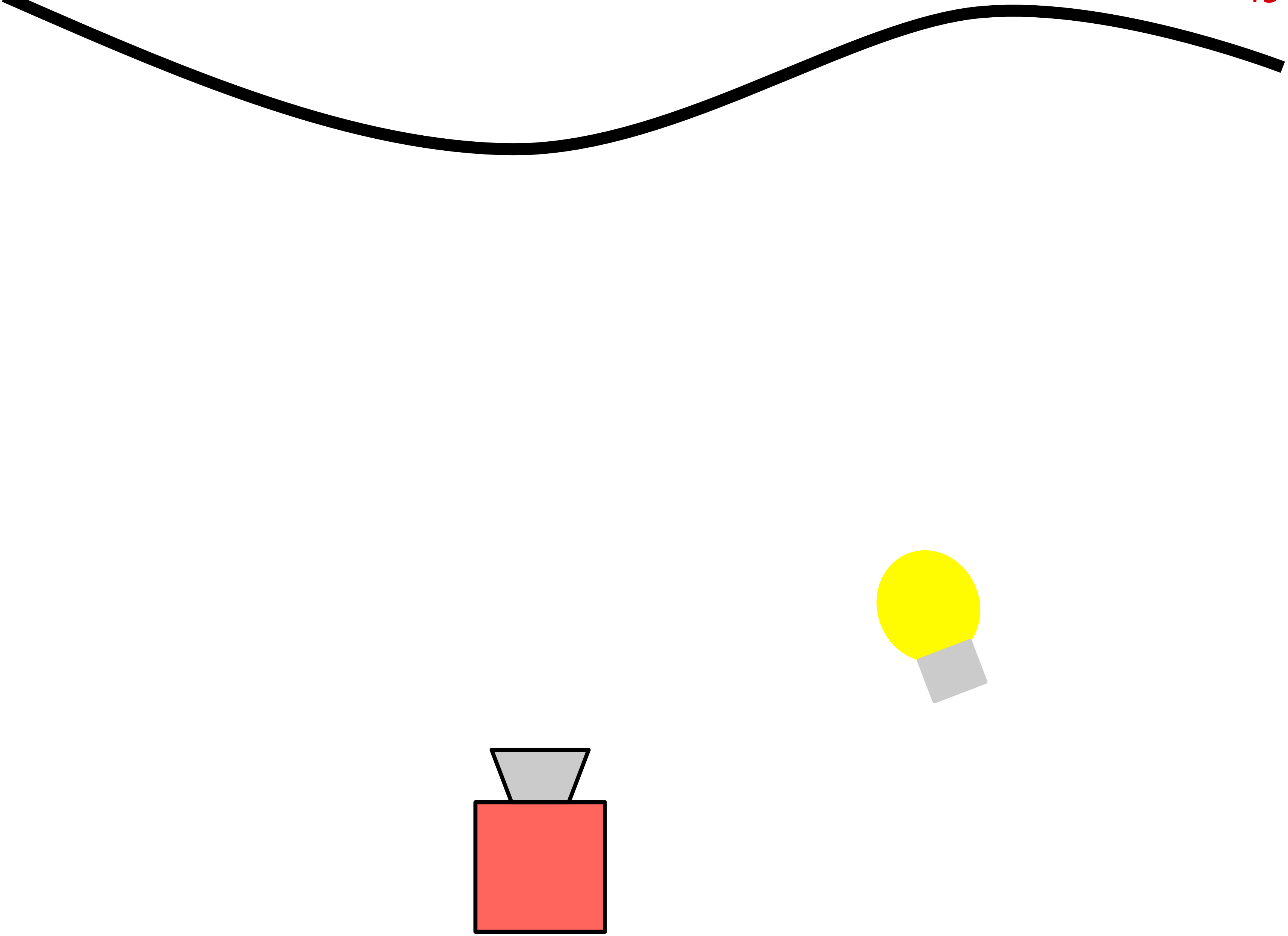


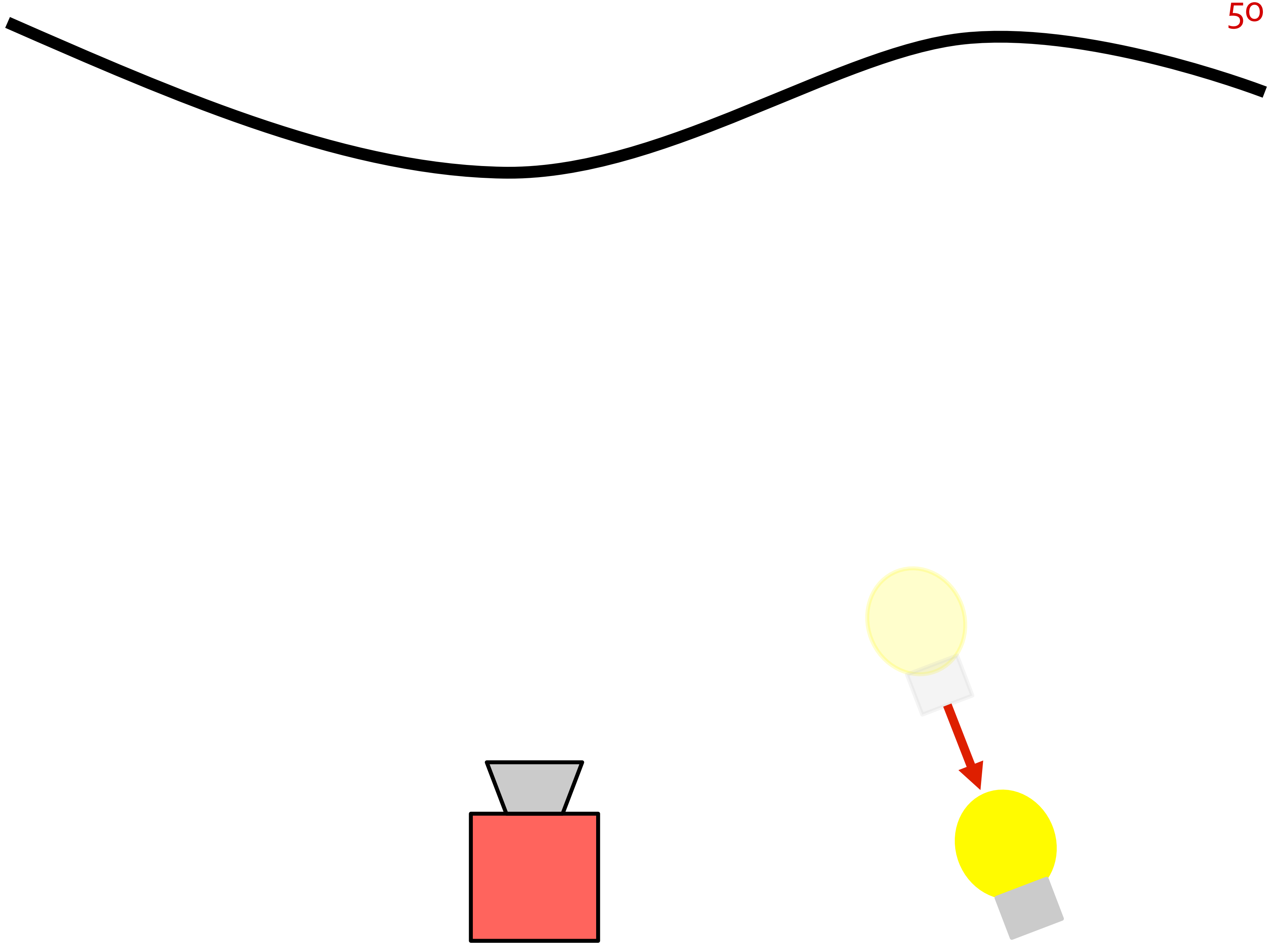




Light Fall-off Stereo

M. Liao, L. Wang, R. Yang, and M. Gong. Light fall-off stereo. In Proceedings of CVPR, pages 1–8, 2007.





Specialized Hardware

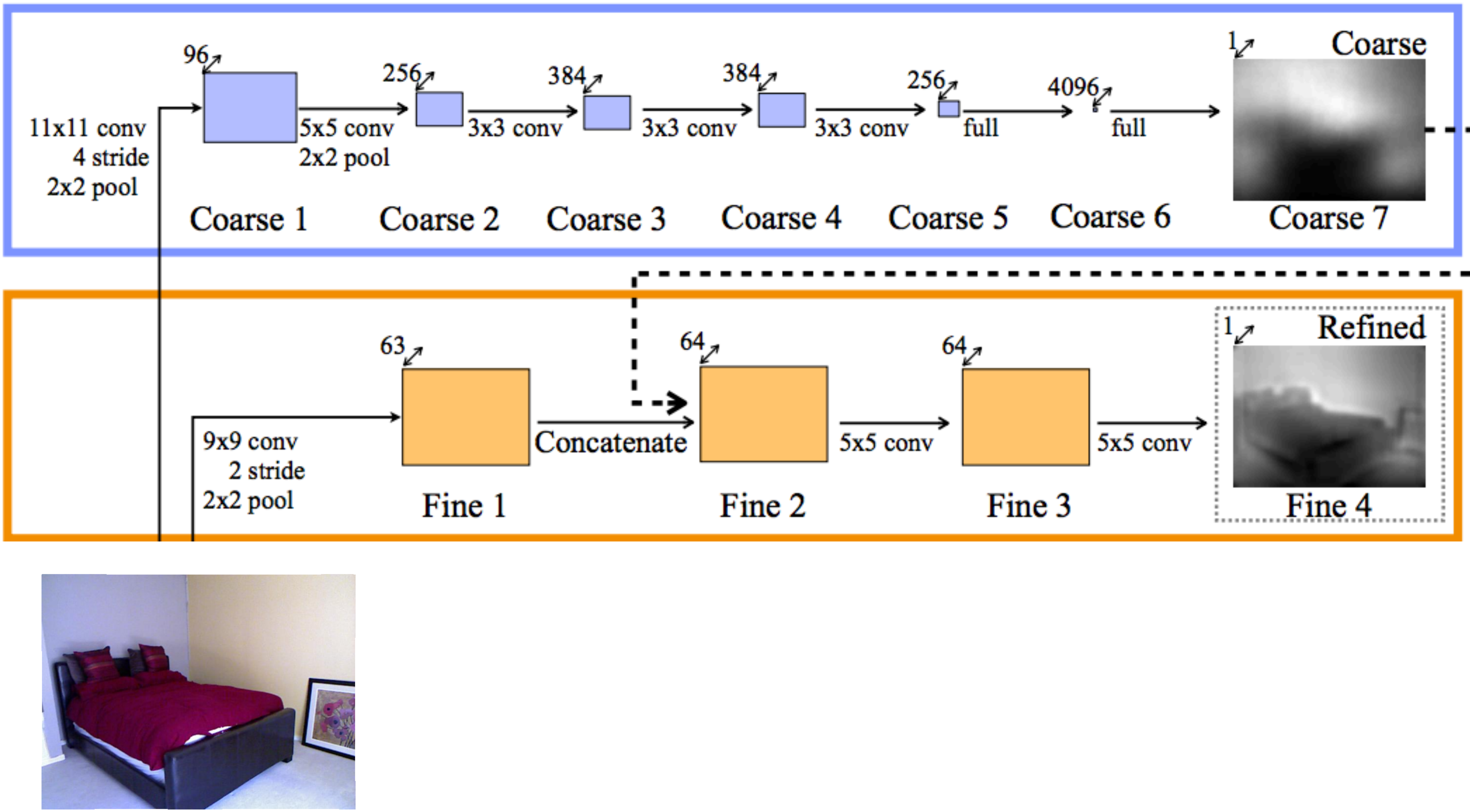
Laser Scanner

Active Illumination

Time of Flight

Estimating Depth

D. Eigen, C. Puhrsch, and R. Fergus. Depth map prediction from a single image using a multi-scale deep network. NIPS 2014


96
256
384
384
256
4096
1
Coarse
11x11 conv
4 stride
2x2 pool
5x5 conv
2x2 pool
3x3 conv
3x3 conv
3x3 conv
full
full
Coarse 1
Coarse 2
Coarse 3
Coarse 4
Coarse 5
Coarse 6
Coarse 7
63
64
64
1
Refined
9x9 conv
2 stride
2x2 pool
Concatenate
5x5 conv
5x5 conv
Fine 1
Fine 2
Fine 3
Fine 4

Train 2 networks

Global coarse-scale network

Local fine-scale network

Global coarse-scale network

Learns a coarse depth map

Used as input to local network

Intuition:

Coarse info learnt already

Focus on learning finer info

Scale ambiguity

Scale invariant error function

$$D(y, y^*) = \frac{1}{2n} \sum_{i=1}^{n} (\log y_i - \log y_i^* + \alpha(y_i, y_i^*))^2$$

$$\alpha(y_i, y_i^*) = \frac{1}{n} \sum_{i=1}^{n} (\log y_i^* - \log y_i)$$

$$D(ay,ay^*) = \frac{1}{2n}\sum_{i=1}^{n}(\log ay_i - \log ay_i^* + \alpha(ay_i,ay_i^*))^2$$

$$D(ay,ay^*) = \frac{1}{2n}\sum_{i=1}^{n}(\log a - \log a + \log y_i - \log y_i^* + \alpha(ay_i,ay_i^*))^2$$

$$D(ay,ay^*) = \frac{1}{2n}\sum_{i=1}^{n}(\log y_i - \log y_i^* + \log a - \log a + \alpha(y_i,y_i^*))^2$$

$$D(ay,ay^*) = D(y,y^*)$$

Loss Function

Scale invariant

$$L(y, y^*) = \frac{1}{n}\sum_{i=1}^{n} d_i^2 - \frac{\lambda}{n^2}(\sum_{i=1}^{n} d_i)^2$$

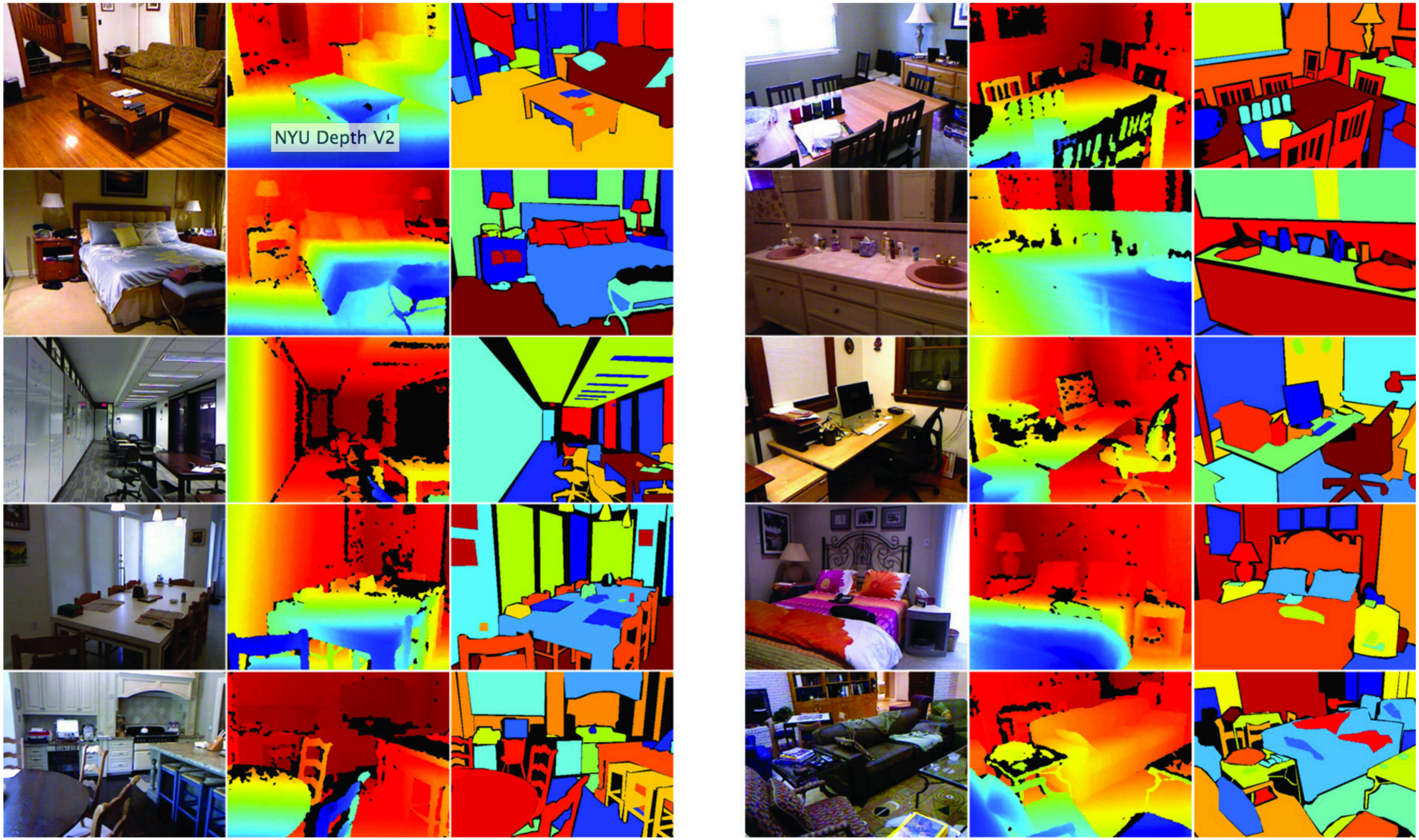
$$d_i = \log y_i - \log y_i^*$$

2 Datasets

NYUDepthV2

N. Silberman, D. Hoiem, P. Kohli, and R. Fergus, "Indoor segmentation and support inference from RGBD images," in *Proc. Eur. Conf. Comput. Vision*, 2012, pp. 746–760.

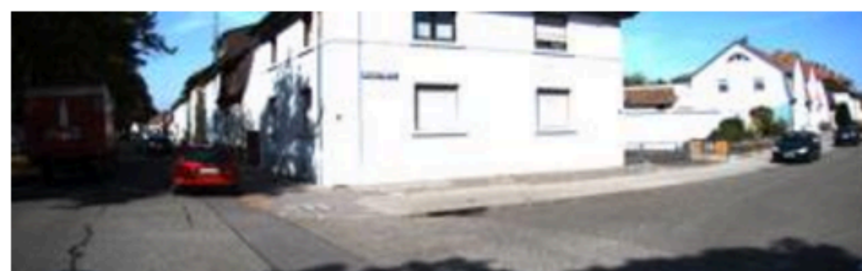
Indoor Rooms


NYU Depth V2

KITTI

A. Greiger, P. Lenz, C. Stiller, and R. Urtasun. Vision meets robotics: The kitti dataset. *International Journal of Robotics Research (IJRR).* 2013.

Outdoor images taken on a car



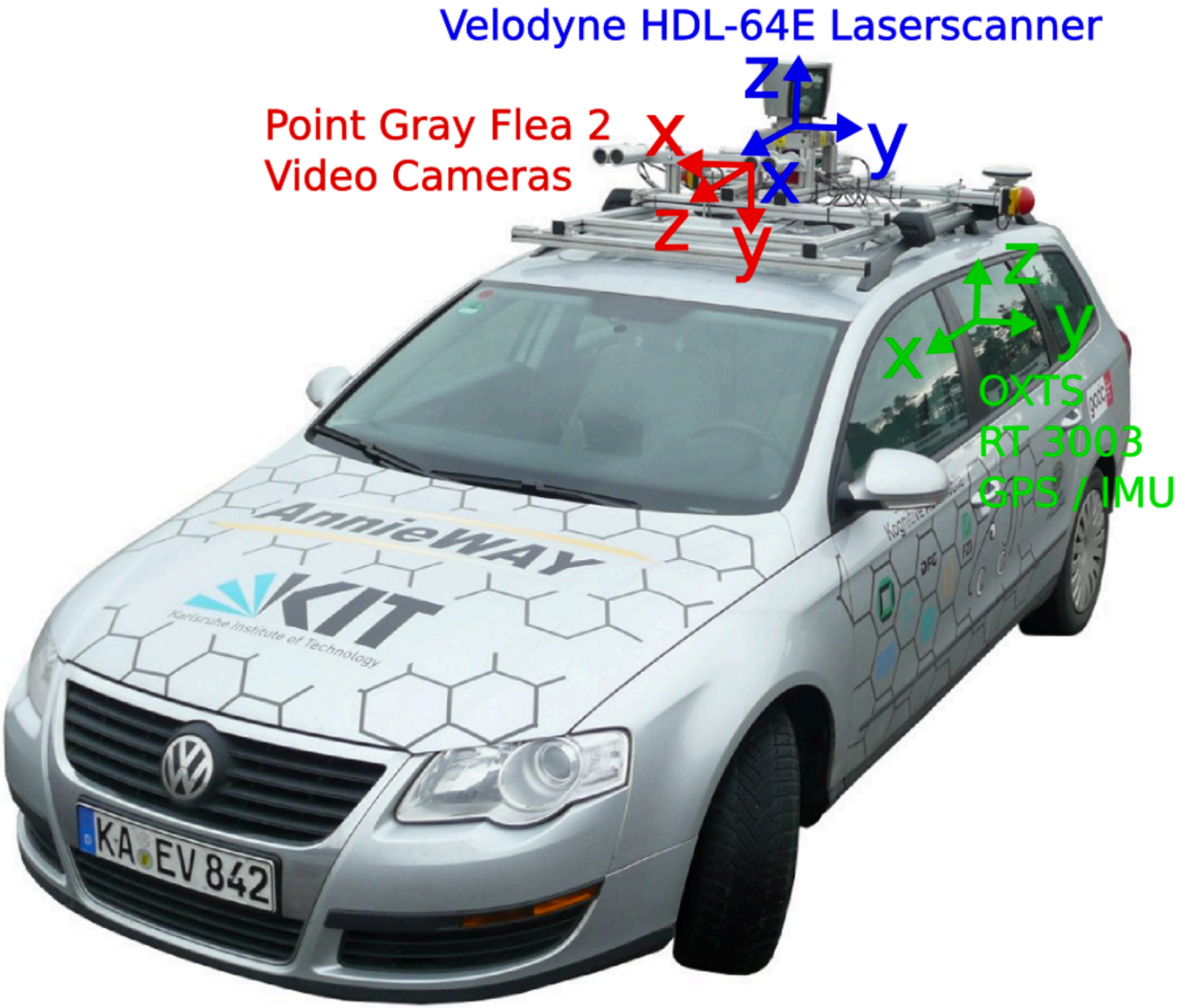
How do you get ground truth?

NYUDepthV2

Kinect


XBOX 360
KINECT

KITTI


Velodyne HDL-64E Laserscanner
z
y
x
Point Gray Flea 2
Video Cameras
x
z
y
z
y
x
OXTS
RT 3003
GPS / IMU
AnnieWAY
KIT
Karlsruhe Institute of Technology
KA EV 842

Time of Flight

Times how long light travels

From light source to camera

Results

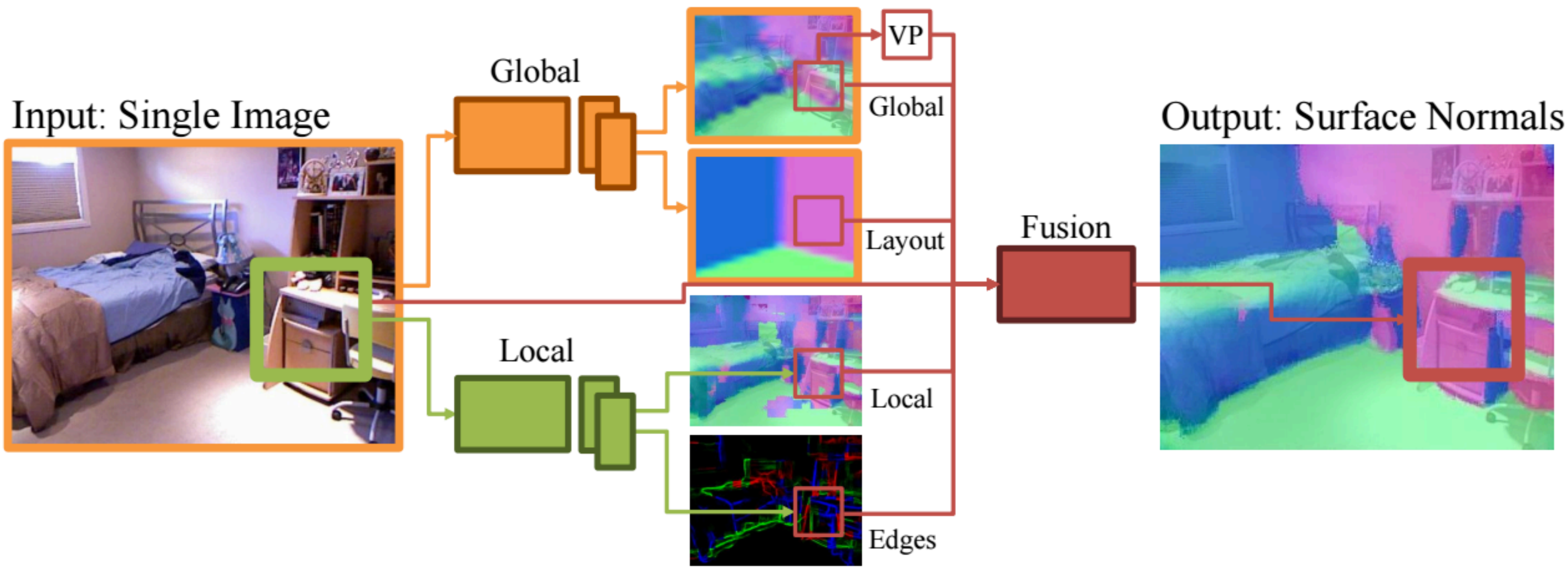
	Mean	Make3D	Ladicky *&al*	Karsch *&al*	Coarse	Coarse + Fine	
threshold $\delta < 1.25$	0.418	0.447	0.542	–	**0.618**	0.611	higher
threshold $\delta < 1.25^2$	0.711	0.745	0.829	–	**0.891**	0.887	is
threshold $\delta < 1.25^3$	0.874	0.897	0.940	–	0.969	**0.971**	better
abs relative difference	0.408	0.349	–	0.350	0.228	**0.215**	
sqr relative difference	0.581	0.492	–	–	0.223	**0.212**	lower
RMSE (linear)	1.244	1.214	–	1.2	**0.871**	0.907	is
RMSE (log)	0.430	0.409	–	–	**0.283**	0.285	better
RMSE (log, scale inv.)	0.304	0.325	–	–	0.221	**0.219**	

Table 1: Comparison on the NYUDepth dataset

Estimating Surface Normals

X. Wang, D. F. Fouhey, and A. Gupta. Designing deep networks for surface normal estimation. CVPR 2015

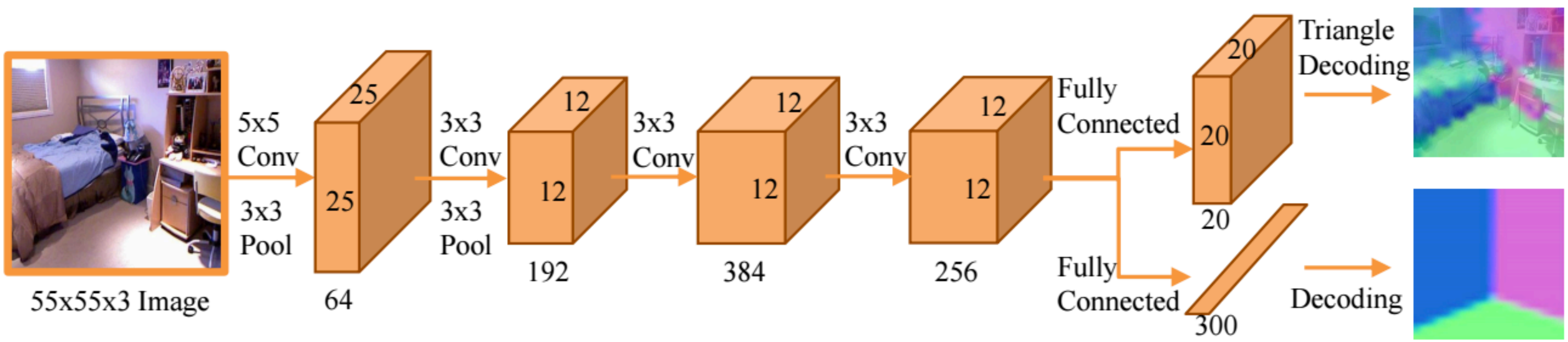
Similar to Eigen



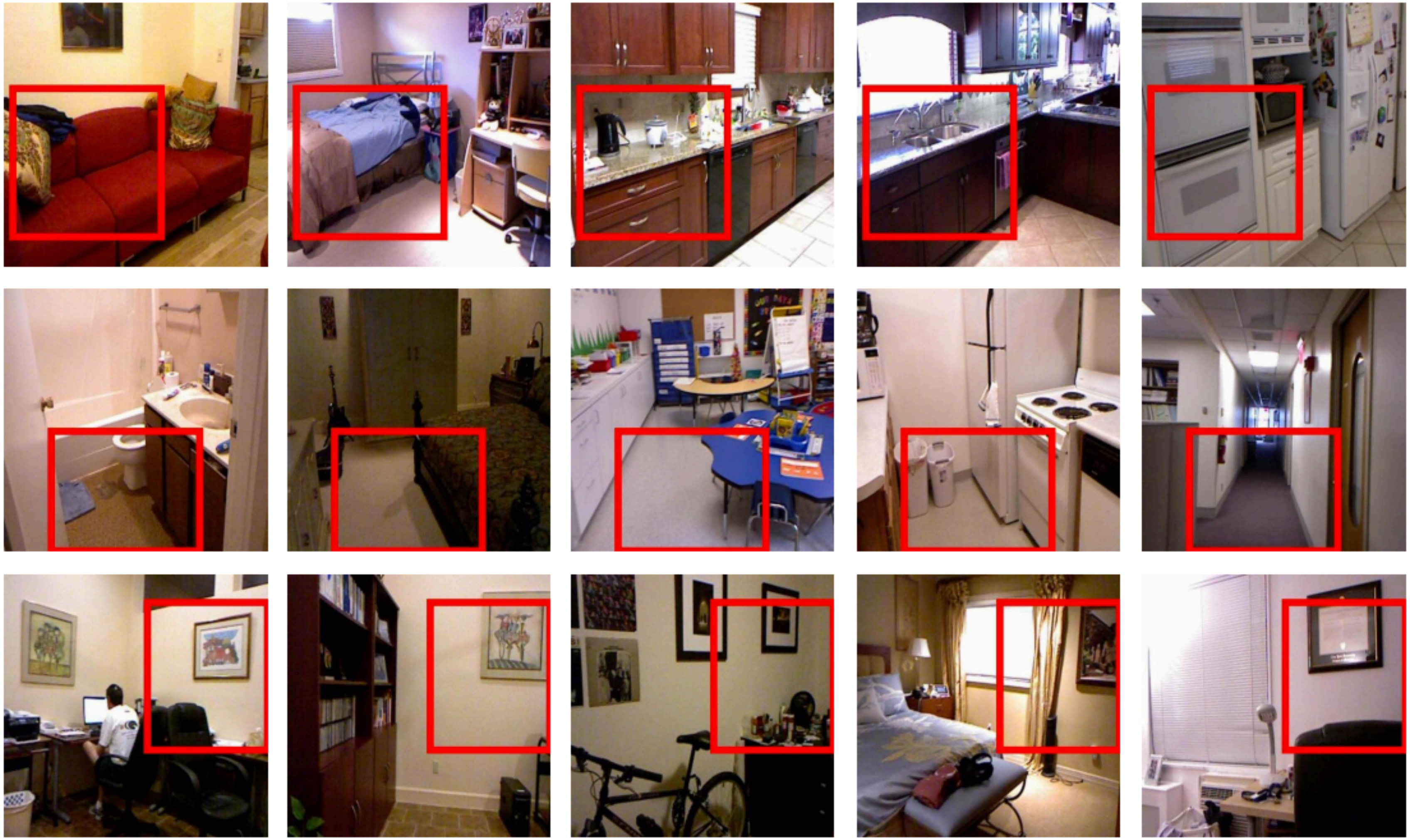
Input: Single Image
Global
Local
VP
Global
Layout
Local
Edges
Fusion
Output: Surface Normals

Trains 3 networks

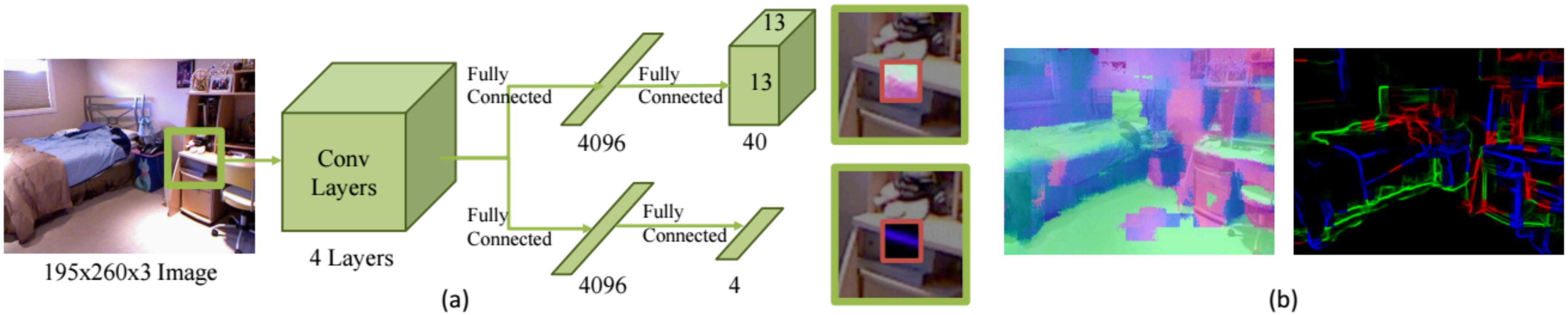
Global coarse-scale network


55x55x3 Image
5x5
Conv
3x3
Pool
25
25
64
3x3
Conv
3x3
Pool
12
12
192
3x3
Conv
12
12
384
3x3
Conv
12
12
256
Fully
Connected
20
20
20
Triangle
Decoding
Fully
Connected
300
Decoding

Trains for room layout as well



Local fine-scale network



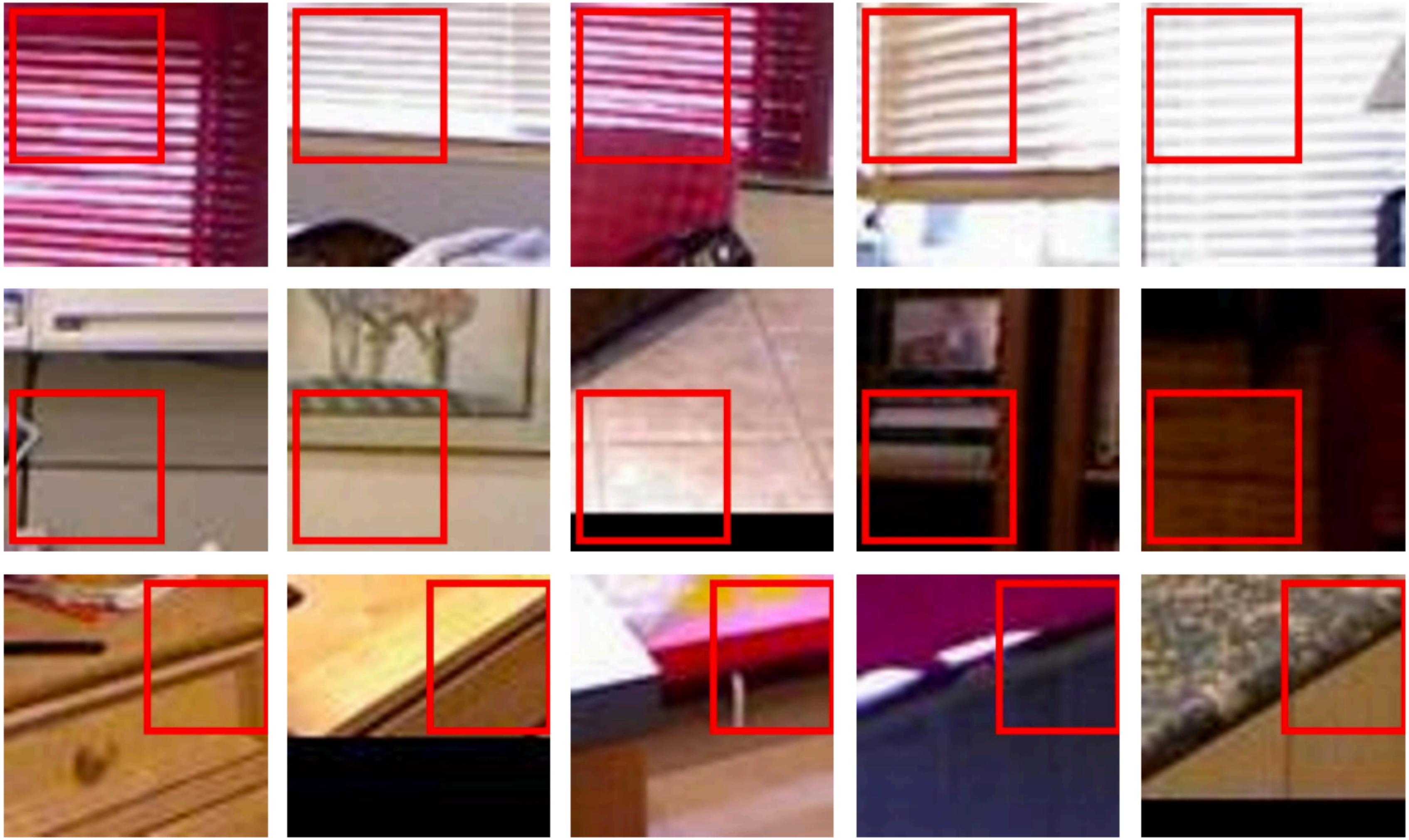
195x260x3 Image
Conv Layers
4 Layers
Fully Connected
4096
Fully Connected
13
13
40
Fully Connected
4096
Fully Connected
4

(a)

(b)

Trains for edge labels as well

Convex, concave, occlusion, N/A



Difference: Global and Local trained separately

Fusion Network

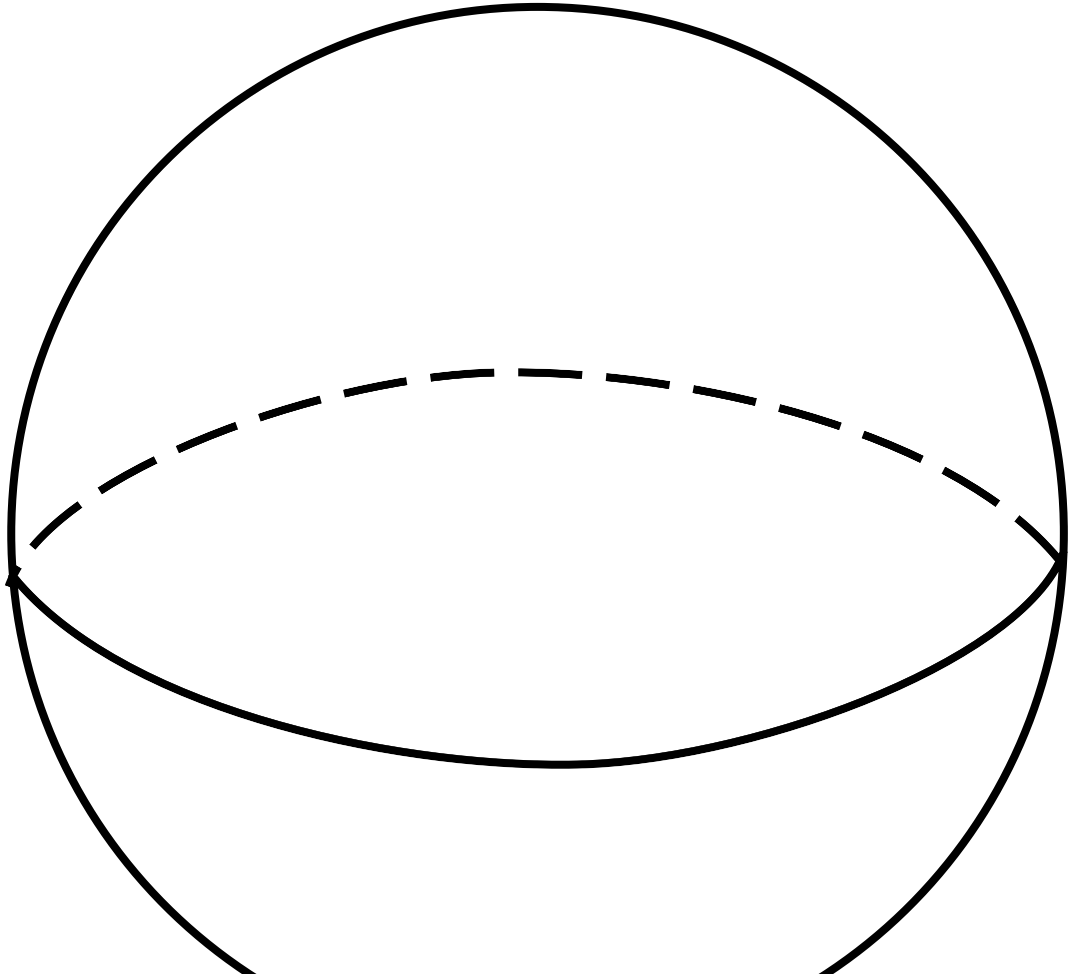
Combines both networks

How to represent normals

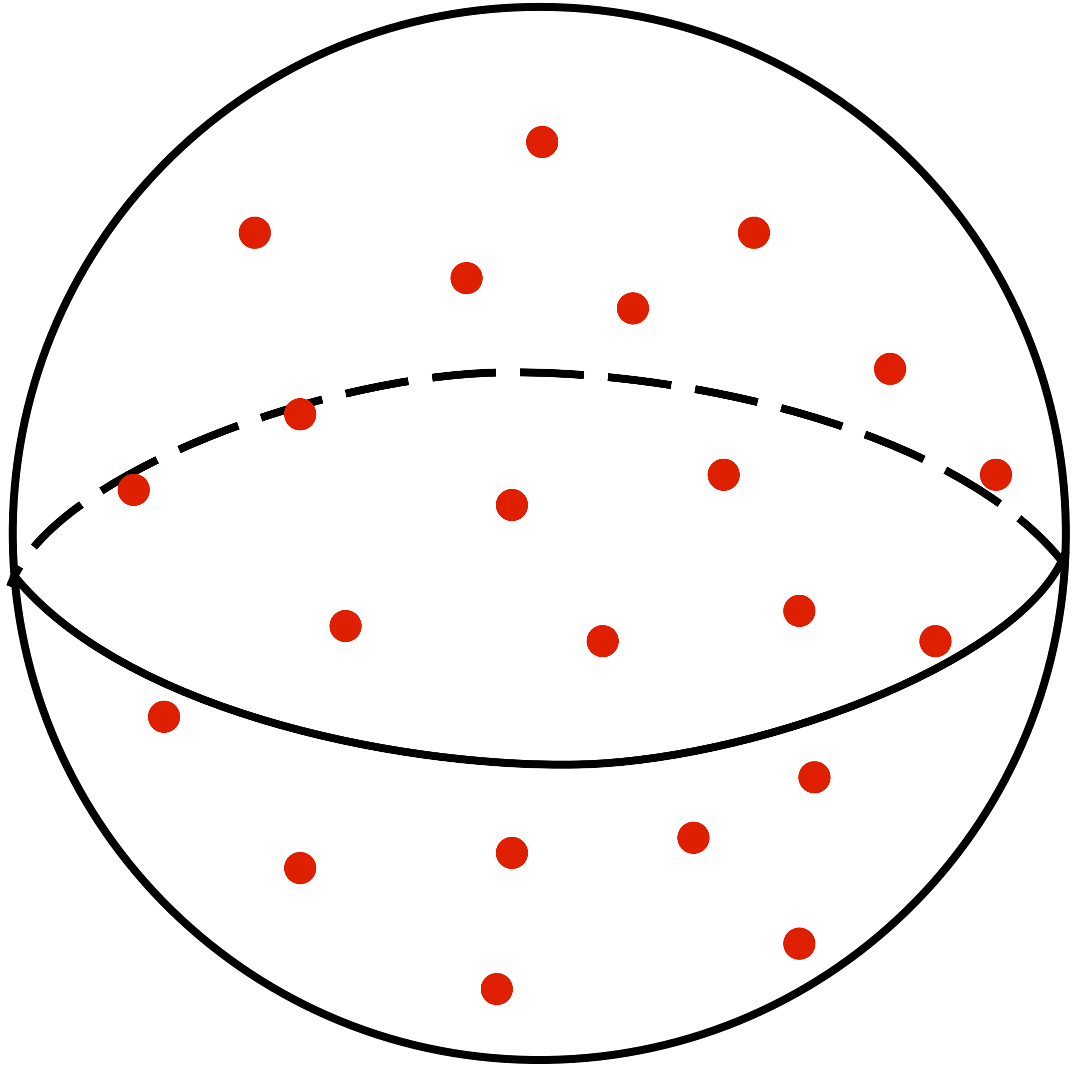
Normals lie in continuous space

Regression as Classification

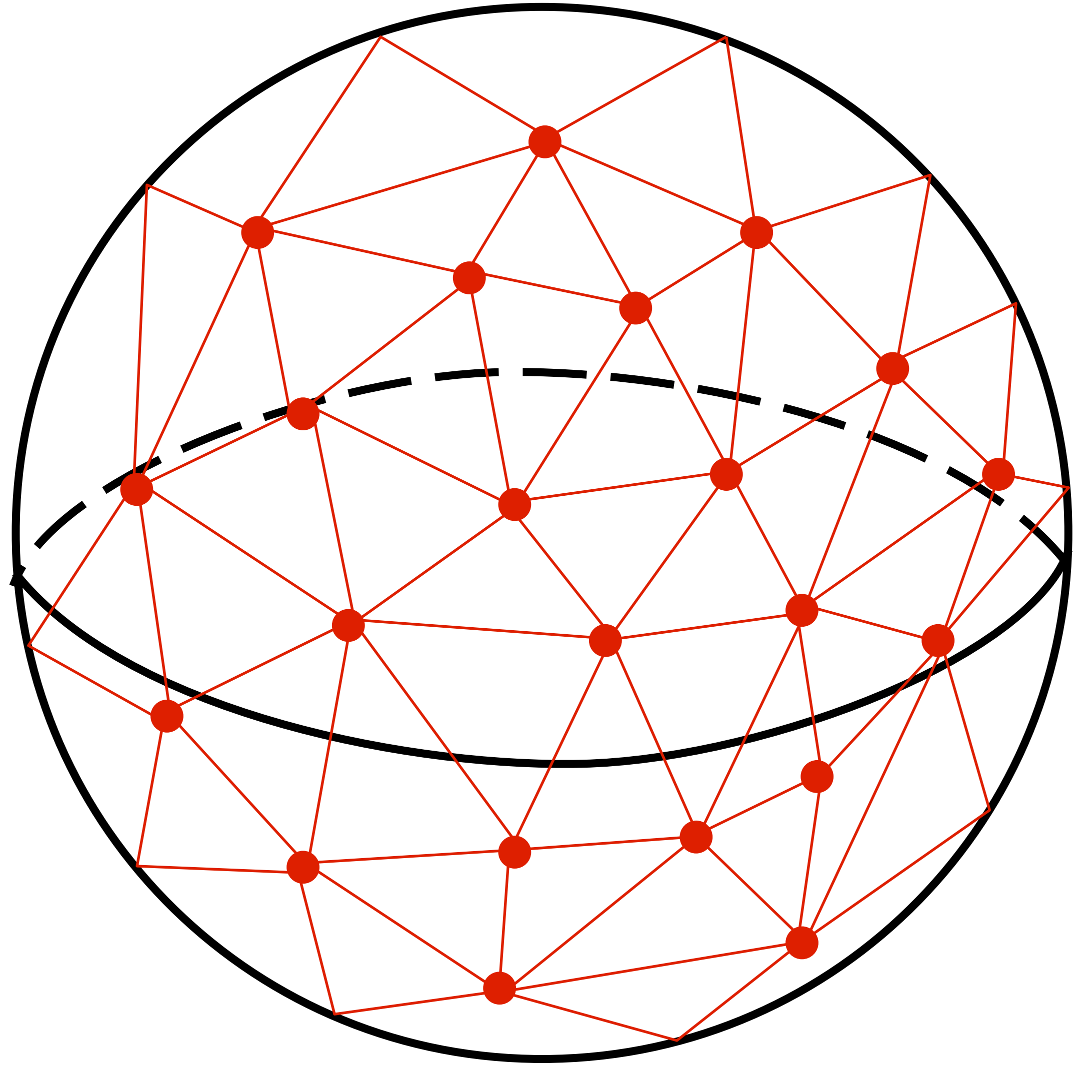
Surface normal triangular coding



Codebook with k-means



Delaunay Triangulation cover



Triangles as classes

Represent Surface Normals

Weighted sum of triangle corners

Loss Function

$$L(I,Y) = -\sum_{i=1}^{M\times M}\sum_{k=1}^{K}(1(y_i = k)\log F_{i,k}(I))$$

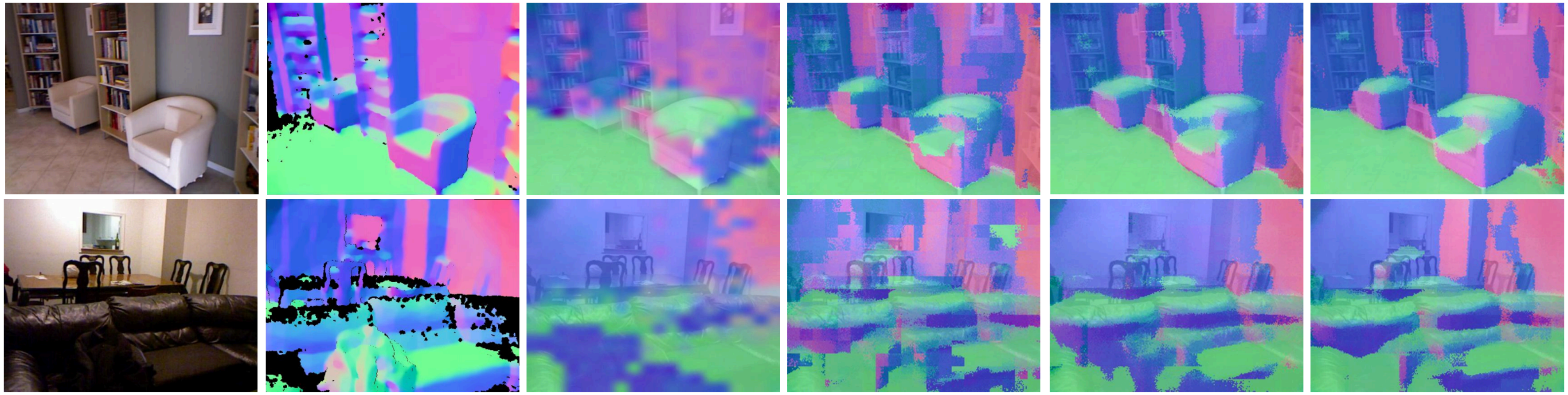


Table 2: Ablative Analysis

	Mean	Median	RMSE	11.25°	22.5°	30°
Full	25.0	**13.8**	35.9	**44.2**	**63.2**	**70.3**
Full (Soft)	**24.2**	17.3	**32.2**	36.8	58.5	68.7
Fusion (+VP)	25.3	14.4	35.9	42.7	62.5	69.9
Fusion (+Edge)	25.8	15.3	36.0	40.0	61.6	69.7
Fusion (+Layout)	25.8	14.9	36.3	41.1	61.9	69.5
Fusion	26.0	15.5	36.2	39.5	61.3	69.3
Bottom-up	32.2	23.5	42.0	27.2	48.5	58.5
Top-down	29.0	19.8	38.3	32.7	53.8	62.4
Eigen et al.(Fusion)	26.8	19.3	35.2	32.6	55.3	65.5
Eigen et al.(Coarse)	27.9	23.4	34.5	25.5	48.4	60.6

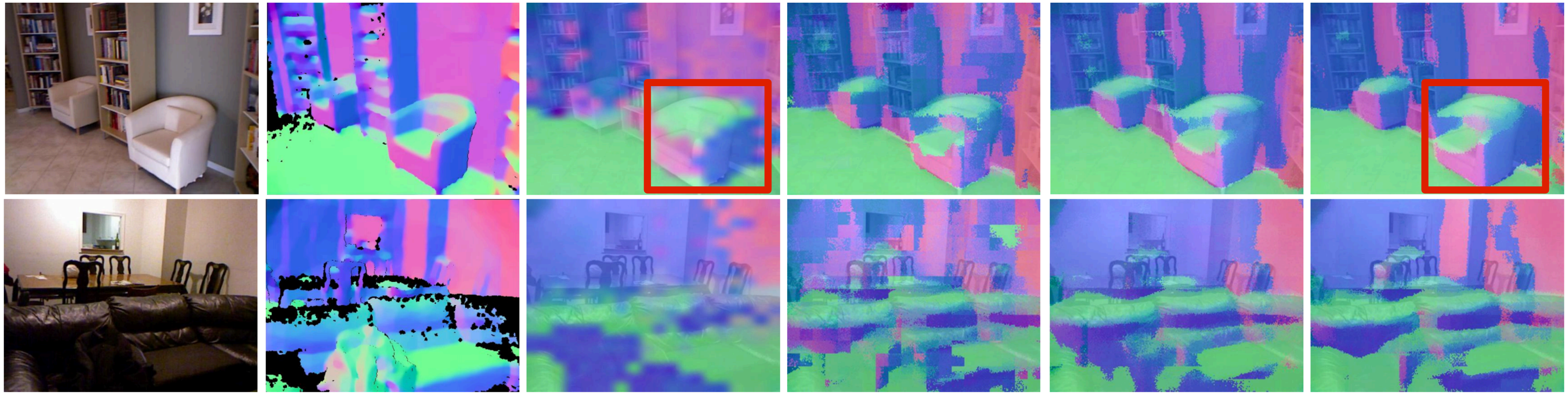


Table 2: Ablative Analysis

	Mean	Median	RMSE	11.25°	22.5°	30°
Full	25.0	**13.8**	35.9	**44.2**	**63.2**	**70.3**
Full (Soft)	**24.2**	17.3	**32.2**	36.8	58.5	68.7
Fusion (+VP)	25.3	14.4	35.9	42.7	62.5	69.9
Fusion (+Edge)	25.8	15.3	36.0	40.0	61.6	69.7
Fusion (+Layout)	25.8	14.9	36.3	41.1	61.9	69.5
Fusion	26.0	15.5	36.2	39.5	61.3	69.3
Bottom-up	32.2	23.5	42.0	27.2	48.5	58.5
Top-down	29.0	19.8	38.3	32.7	53.8	62.4
Eigen et al.(Fusion)	26.8	19.3	35.2	32.6	55.3	65.5
Eigen et al.(Coarse)	27.9	23.4	34.5	25.5	48.4	60.6

Thoughts

Do not address bas-relief

Incorporate Computer Graphics

Inverse problem

Given surface normals

How should the scene look?

What is the correct image?

Incorporate image formation model

Why depth from single image