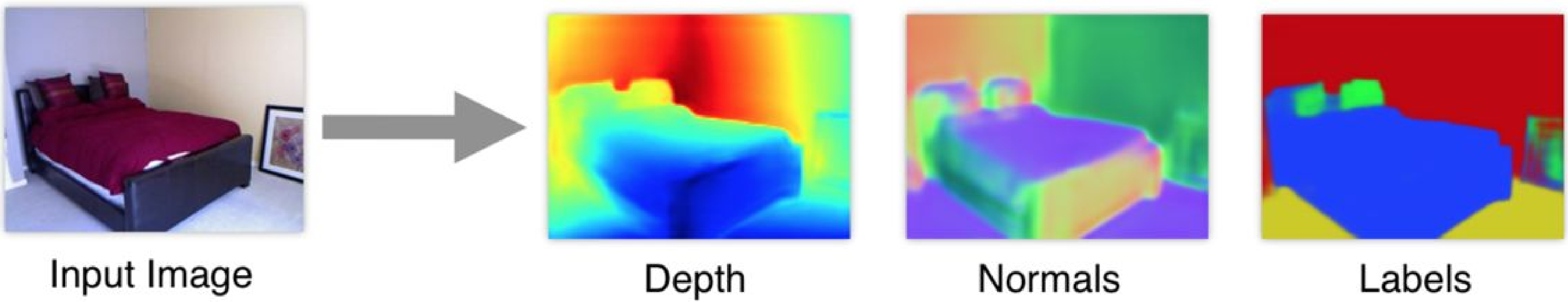
Predicting Depth, Surface Normals and Semantic Labels with a Common Multi-Scale Convolutional Architecture

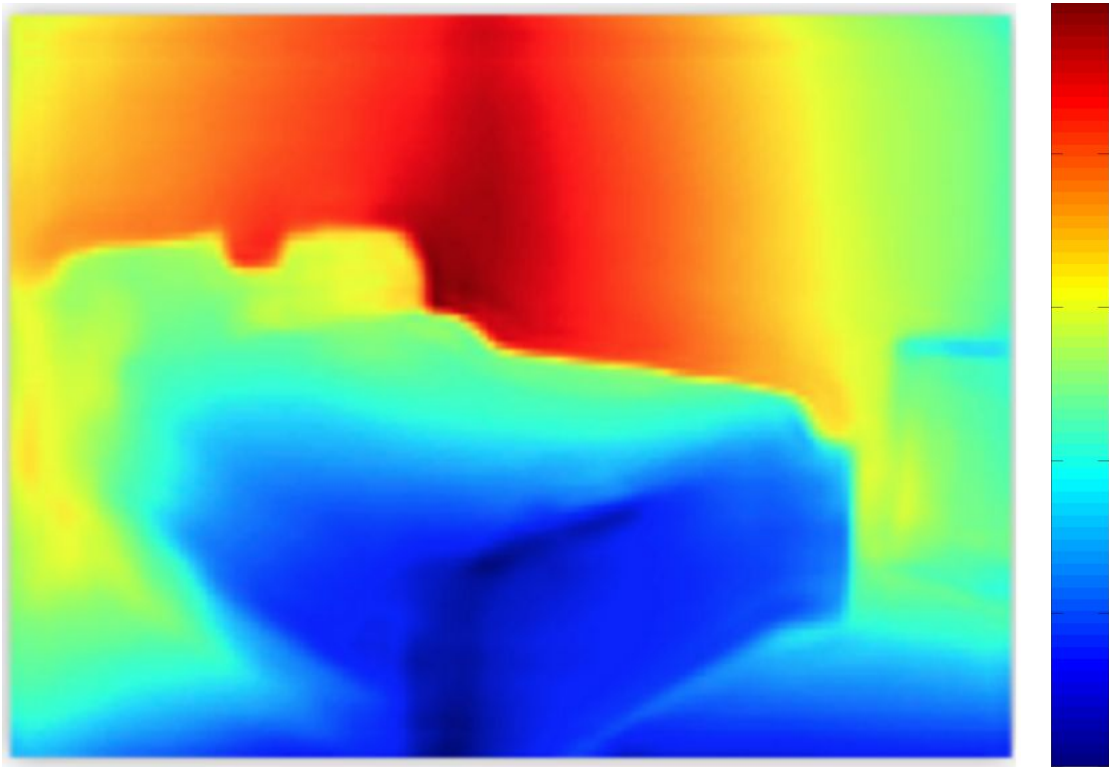
David Eigen, Rob Fergus

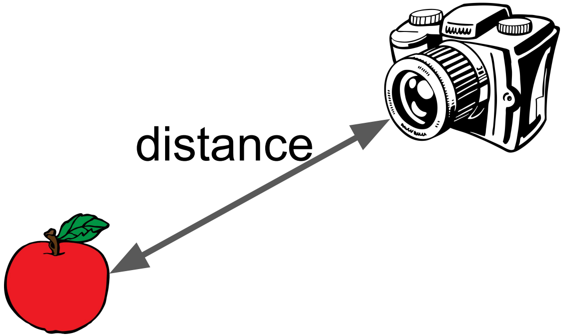


Presented by: Rex Ying and Charles Qi

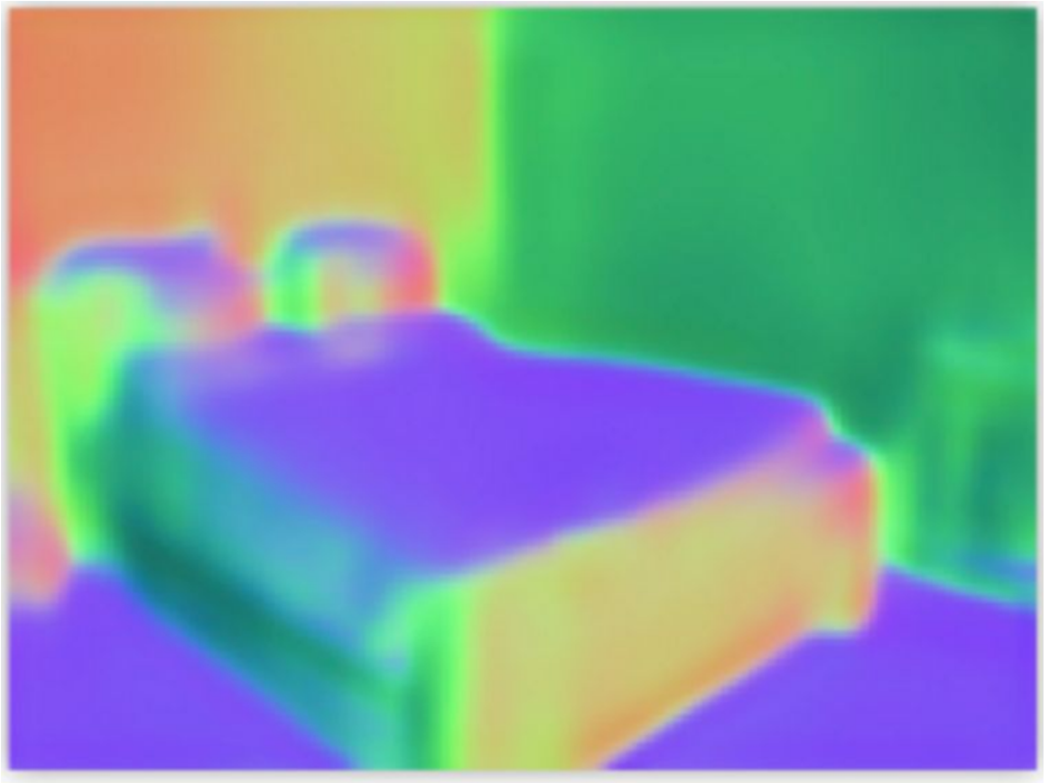
Input: A Single RGB Image

Estimate Depth



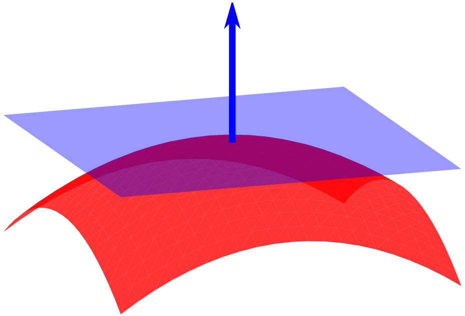


Estimate Surface Normals

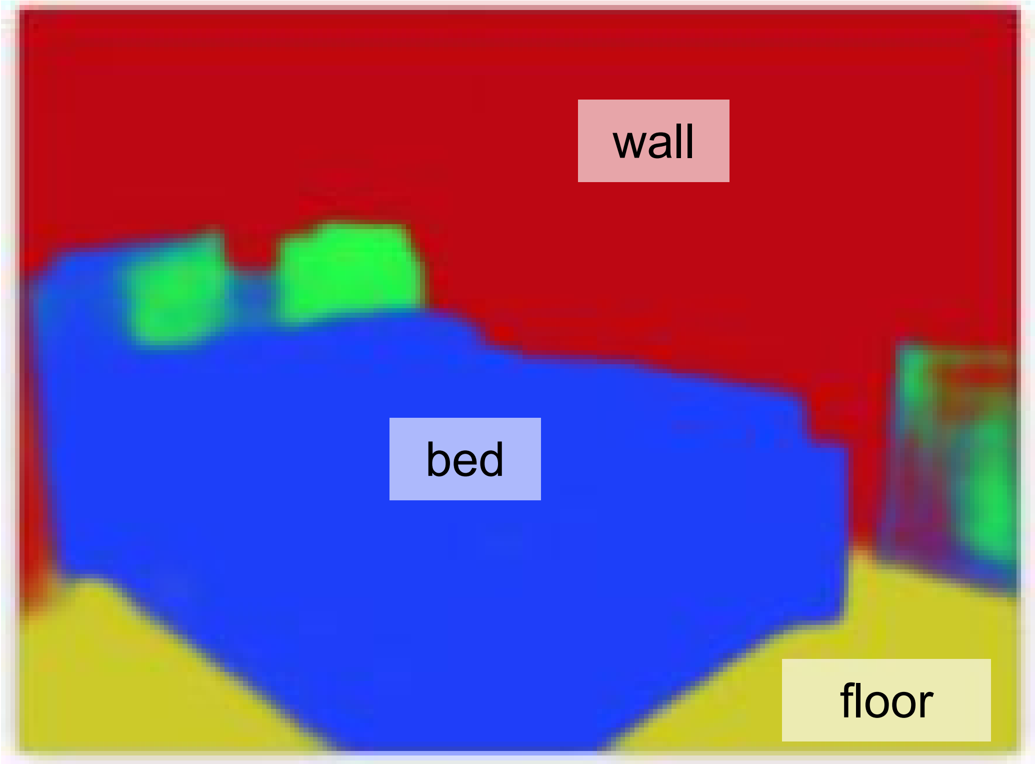


(X, Y, Z) normal vector

Surface normal at point P is a vector that is perpendicular to the tangent plane to that surface at P.

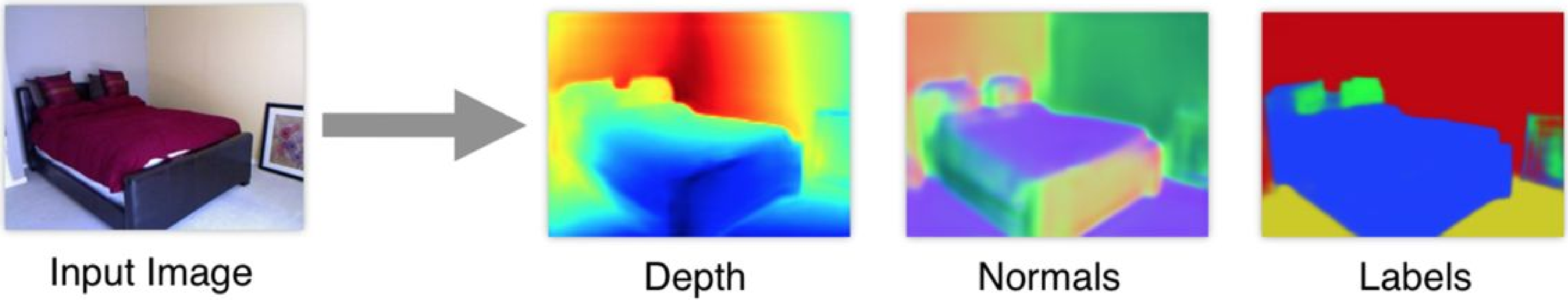


Predict Per-pixel Semantic Labels



A label L is assigned to each pixel, indicating which category (bed, pillow, wall, floor etc.) this pixel belongs to.

3D Scene Representation = Geometry + Semantics

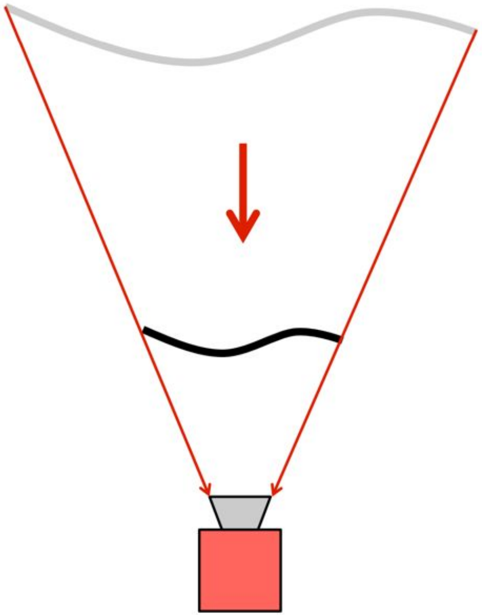


Representing 3D space with depth map and orientation map

Physical Geometry

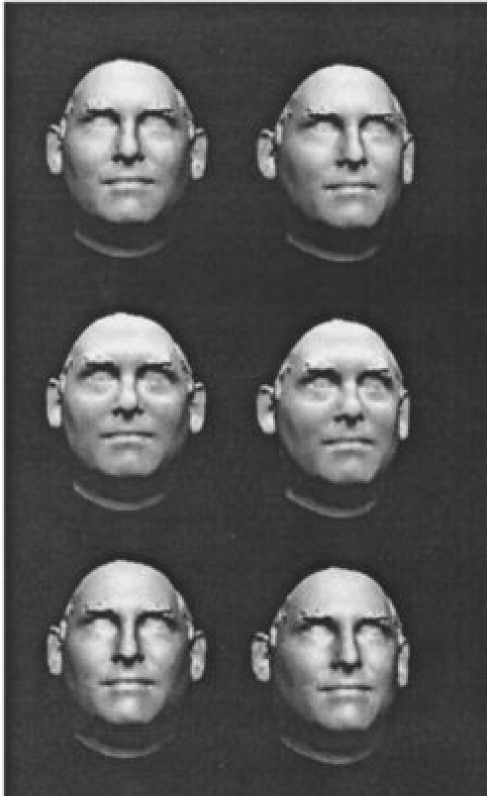
High-level Semantics

Predicting 3D Geometry is Hard: Multiple Ambiguities!

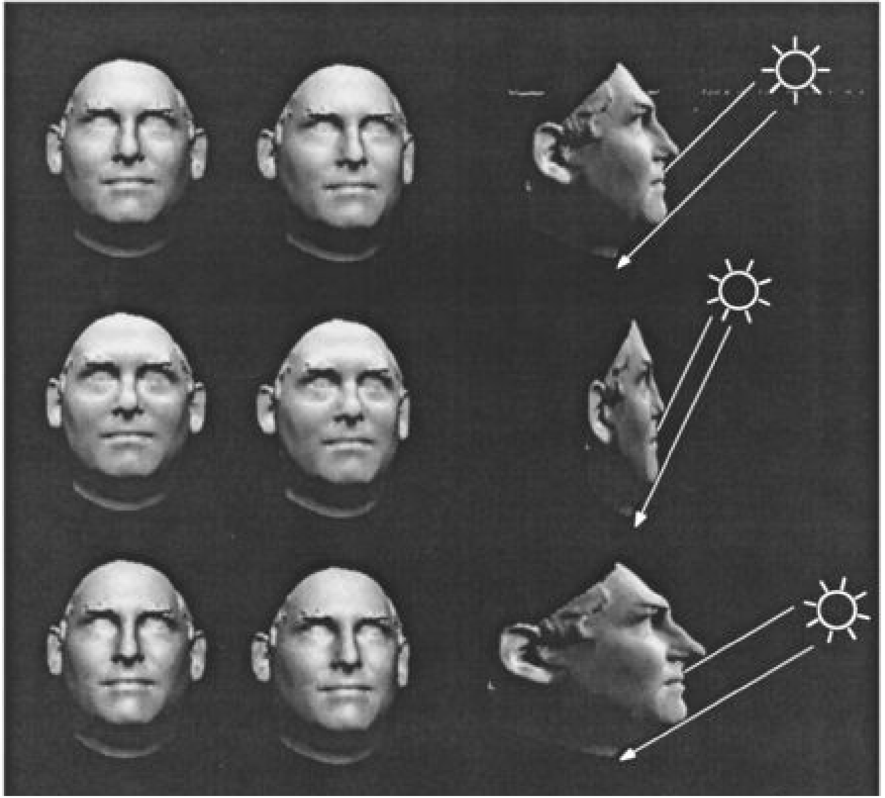


Scale Ambiguity

Predicting 3D Geometry is Hard: Multiple Ambiguities!



Predicting 3D Geometry is Hard: Multiple Ambiguities!



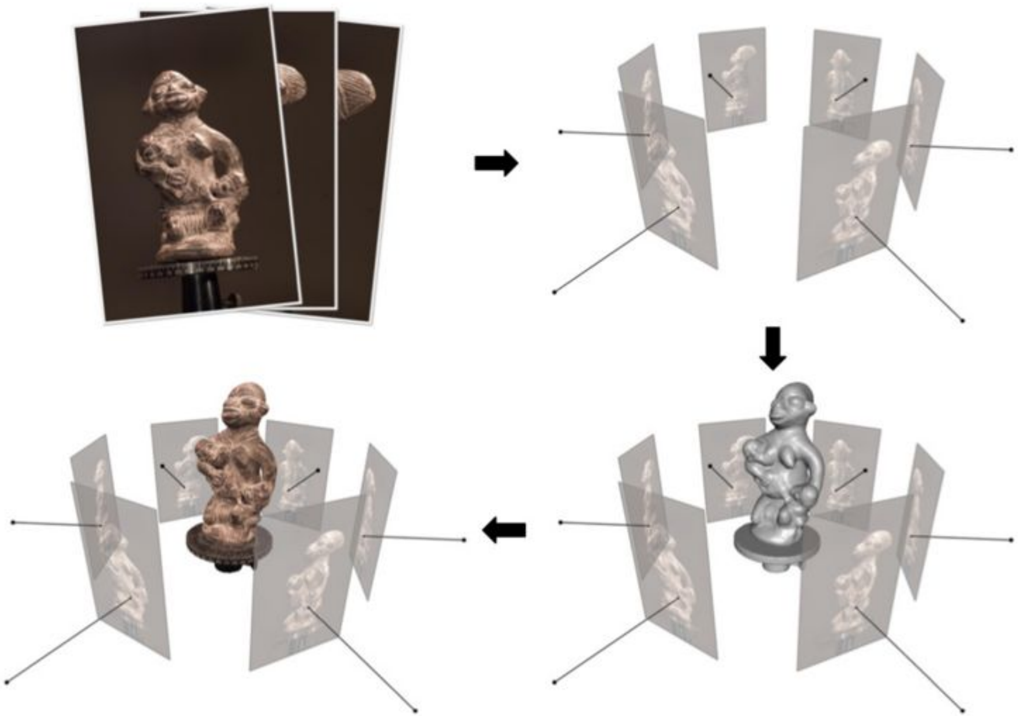
Bas-Relief Ambiguity

Belhumeur, Peter N., David J. Kriegman, and Alan L. Yuille. "The bas-relief ambiguity." International journal of computer vision 35.1 (1999): 33-44.

How to acquire 3D geometry of a scene?

Traditional Methods

Multi-view Stereo

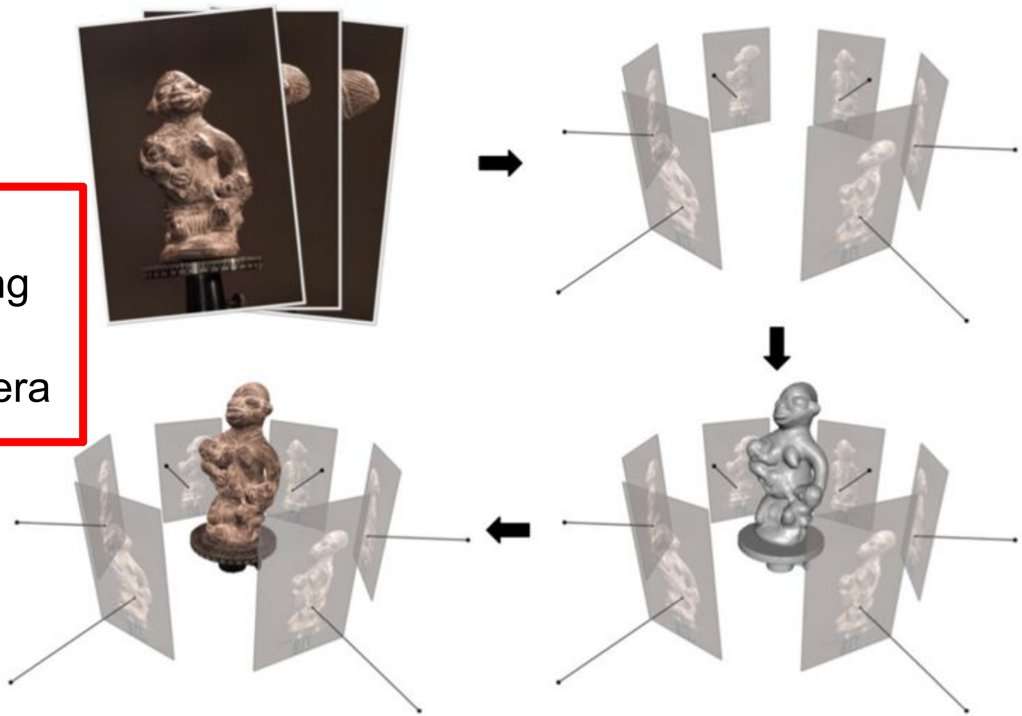


Multi-View Stereo: A Tutorial - Carlos Hernandez, 2013

Traditional Methods

Multi-view Stereo

- \+ More deterministic than ConvNet
- \- Finding correspondence is challenging
- \- Require multiple images input
- \- (Usually) require well calibrated camera

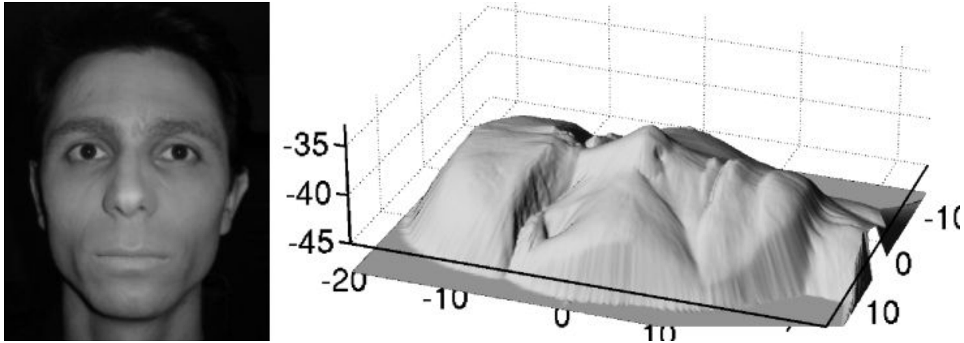


Multi-View Stereo: A Tutorial - Carlos Hernandez, 2013

Traditional Methods

Shape from X

- X =Shading
- X =Multiple Light Sources (photometric stereo)
- X =Texture
- X =Focus/Defocus
- X =Specularities
- X =Shadows

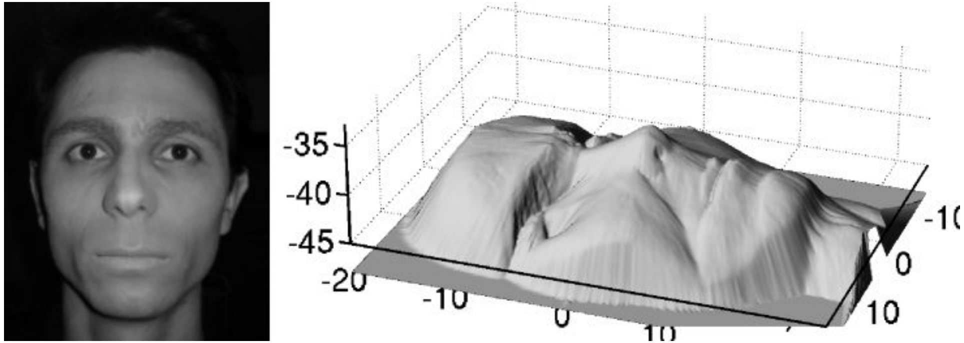


Prados, Emmanuel, and Olivier Faugeras. "Shape from shading: a well-posed problem?." CVPR 2005

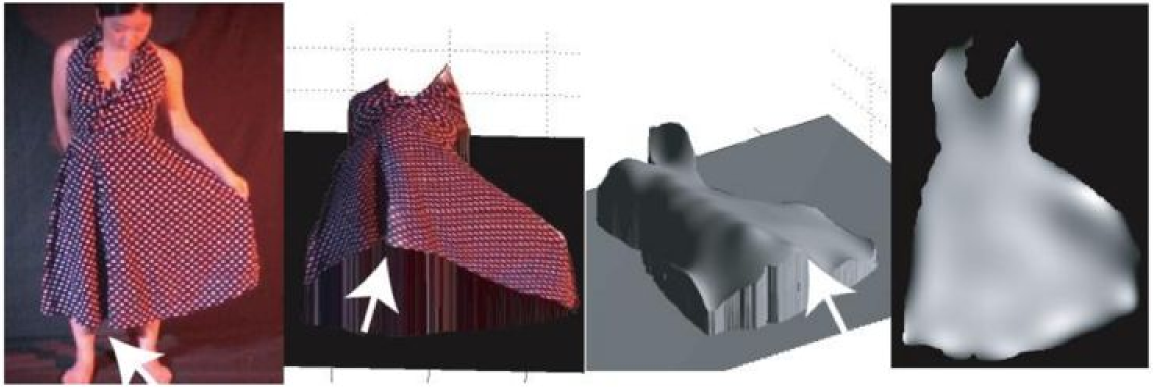
Traditional Methods

Shape from X

- X =Shading
- X =Multiple Light Sources (photometric stereo)
- X =Texture
- X =Focus/Defocus
- X =Specularities
- X =Shadows



Prados, Emmanuel, and Olivier Faugeras. "Shape from shading: a well-posed problem?." CVPR 2005

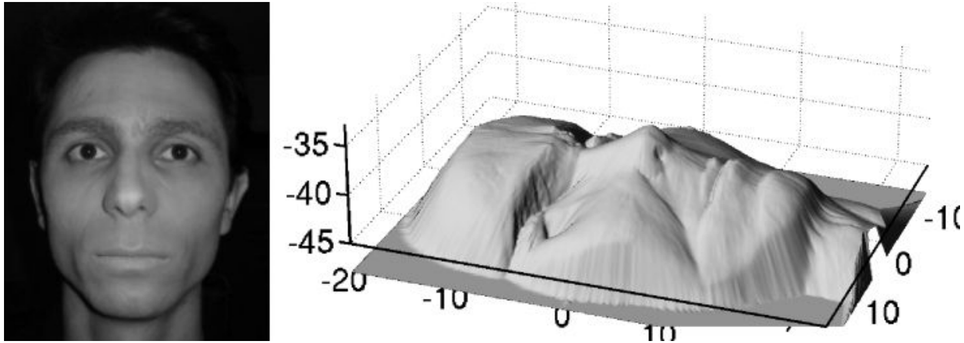


Lobay, Anthony, and David A. Forsyth. "Recovering shape and irradiance maps from rich dense texton fields." *CVPR 2004*

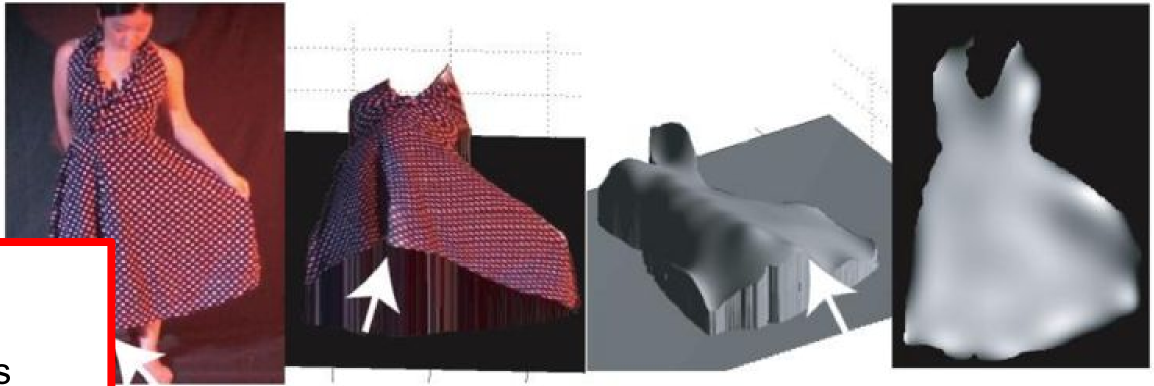
Traditional Methods

Shape from X

X =Shading
X =Multiple Light Sources (photometric stereo)
X =Texture
X =Focus/Defocus
X =Specularities
X =Shadows



Prados, Emmanuel, and Olivier Faugeras. "Shape from shading: a well-posed problem?." CVPR 2005



- \+ Relatively accurate geometry
- \- Lots of assumptions
- \- Not generalizing well to scenes

ony, and David A. Forsyth. "Recovering shape and irradiance maps from rich dense texton fields." *CVPR 2004*

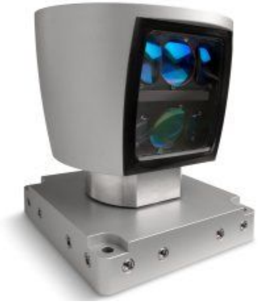
Traditional Methods

Specialized Hardware

- \+ Ground truth depth
- \- Cost of data acquisition and HW

Structured Light

Time of Flight



Laser Scanner

Previous Methods

Make3D

Make3D: Learning 3-D Scene Structure from a Single Still Image, Ashutosh Saxena, Min Sun, Andrew Y. Ng, In IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI), 2008.

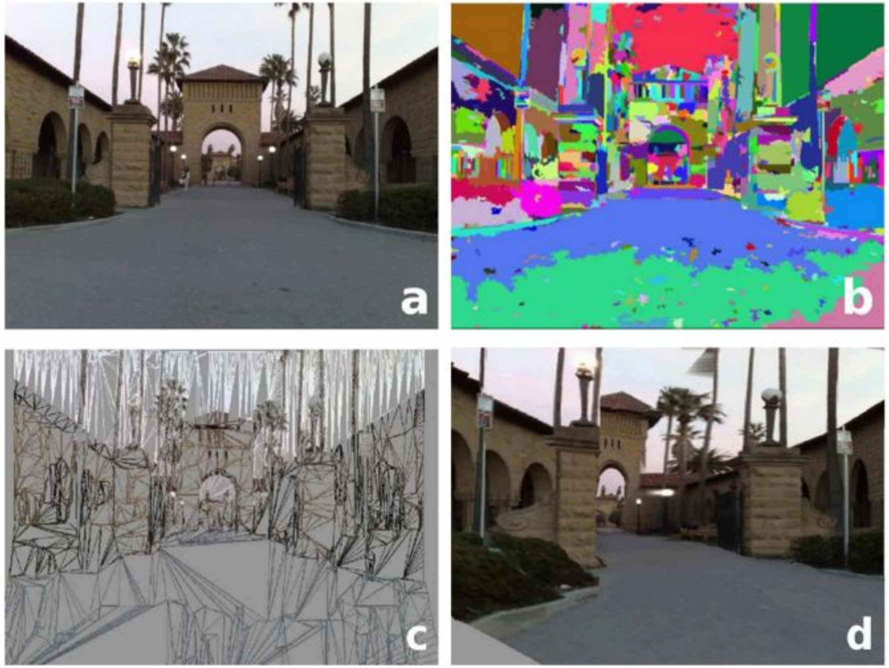


Fig. 1. (a) An original image. (b) Oversegmentation of the image to obtain "superpixels". (c) The 3-d model predicted by the algorithm. (d) A screenshot of the textured 3-d model.

Previous Methods

Make3D

Make3D: Learning 3-D Scene Structure from a Single Still Image, Ashutosh Saxena, Min Sun, Andrew Y. Ng, In IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI), 2008.

+ Flexible, input is a single image
+ Good pioneer work (showing the feasibility of the problem)
- Hand crafted pipeline

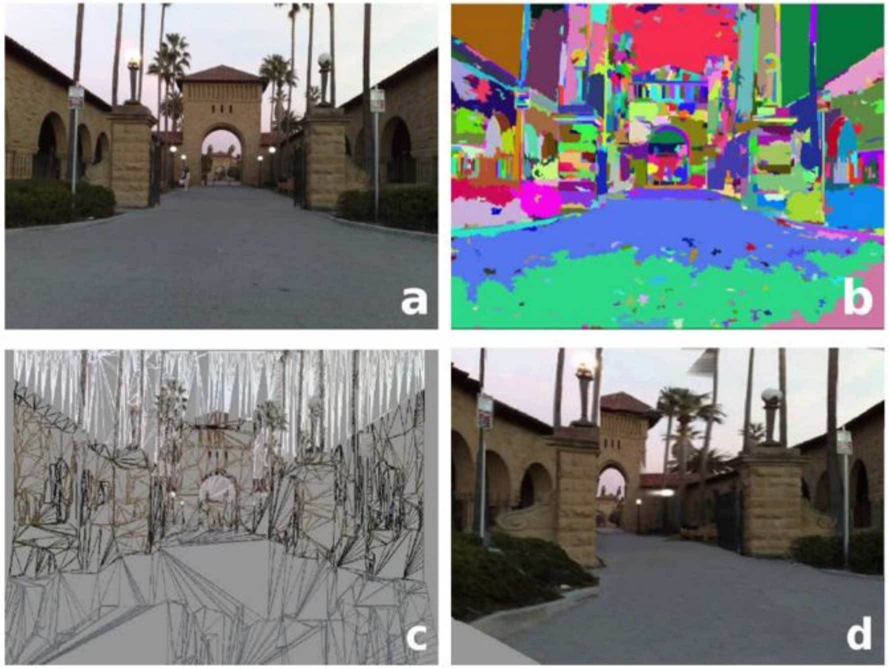
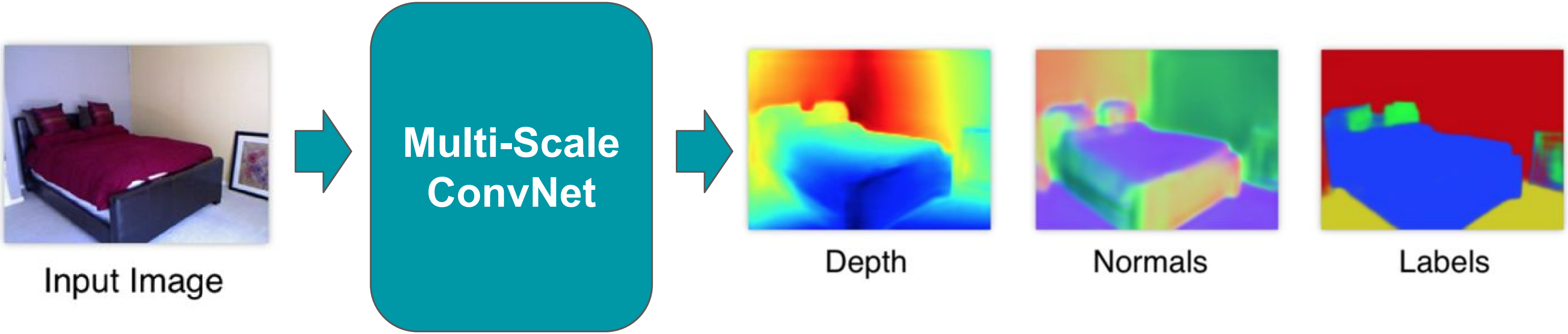


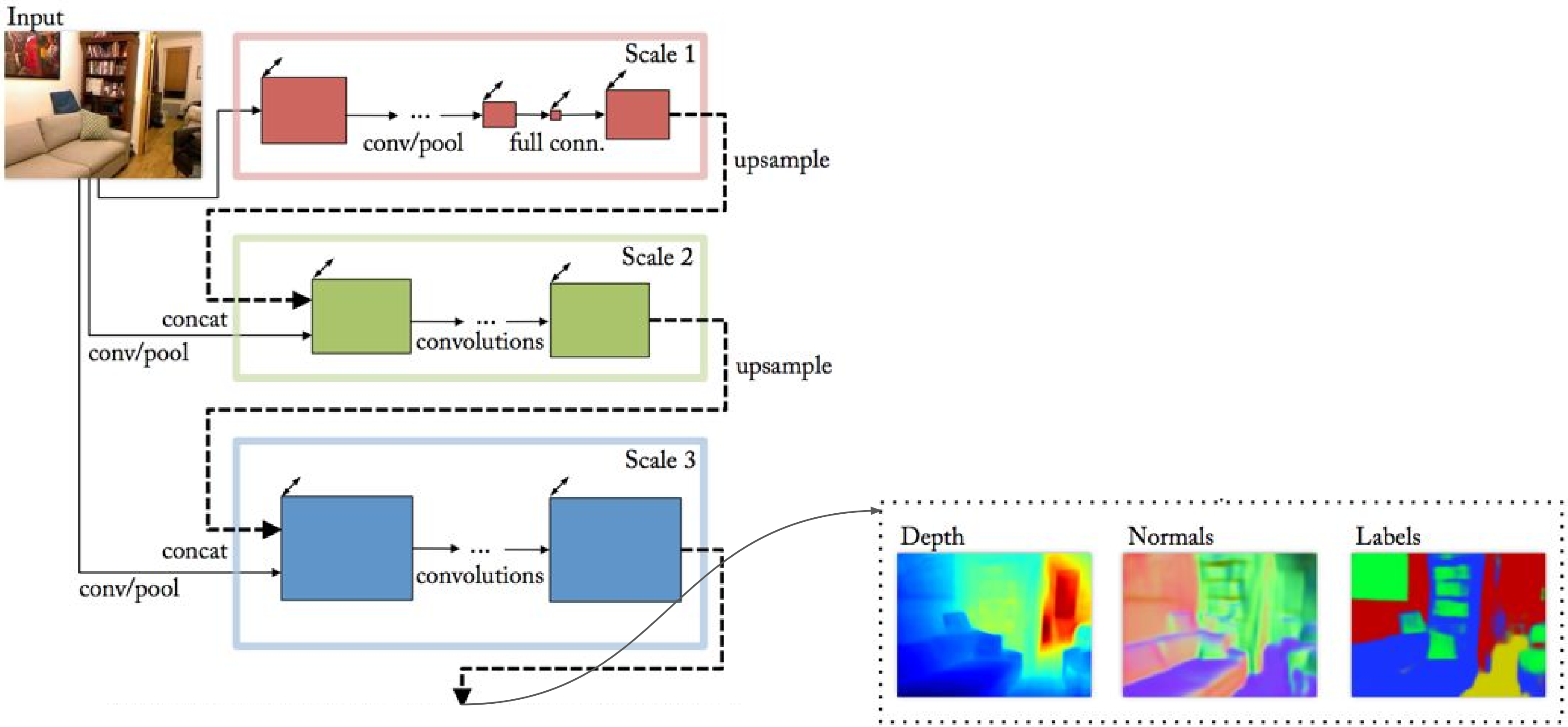
Fig. 1. (a) An original image. (b) Oversegmentation of the image to obtain “superpixels”. (c) The 3-d model predicted by the algorithm. (d) A screenshot of the textured 3-d model.

This Paper

Three Tasks in a Uniform Framework
End-to-End Learning



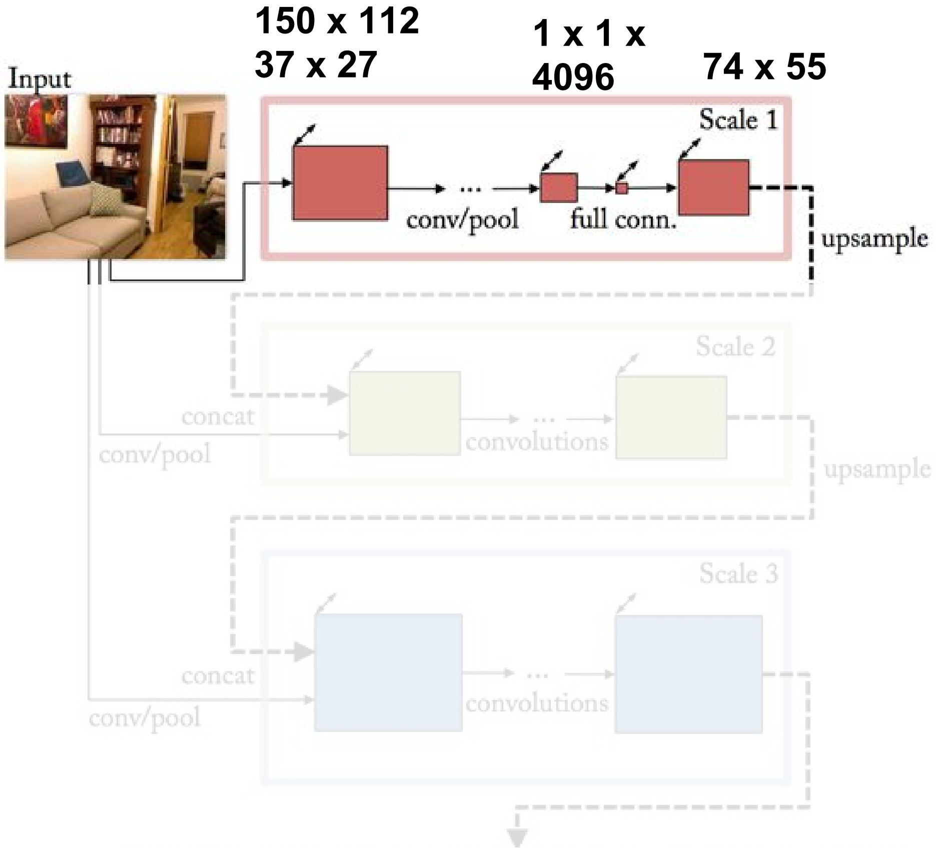
Coarse To Fine Multi-Scale ConvNet



Multiscale Architecture

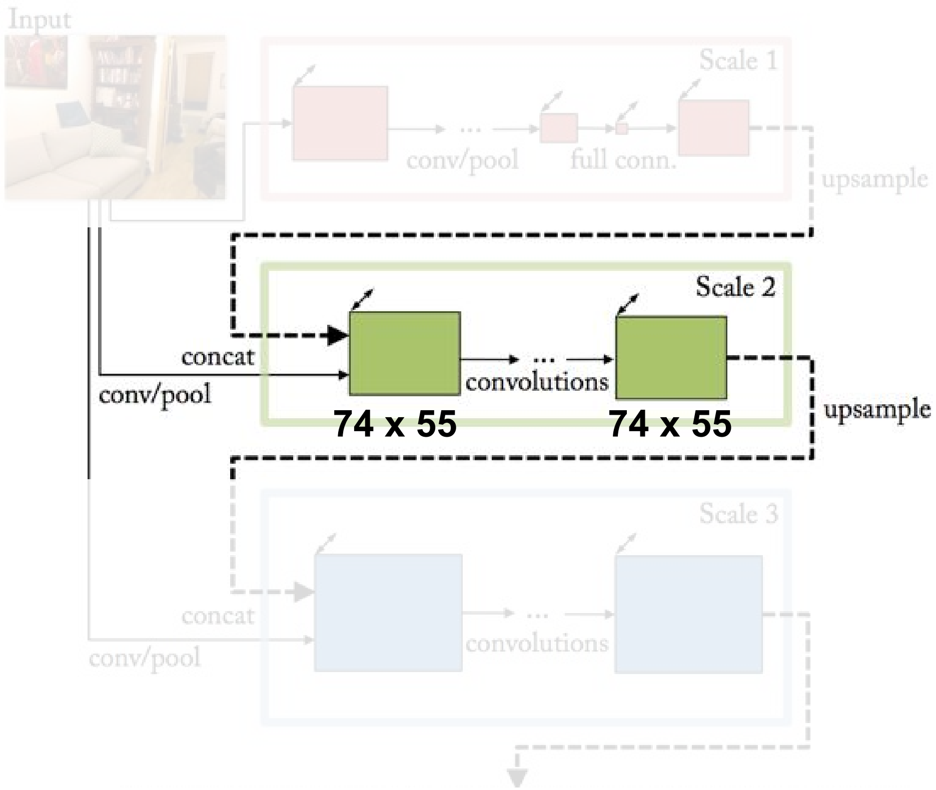
- Networks corresponding to different scales are connected in series
- From low resolution to high resolution
- Can naturally be used to perform many different tasks

Scale 1



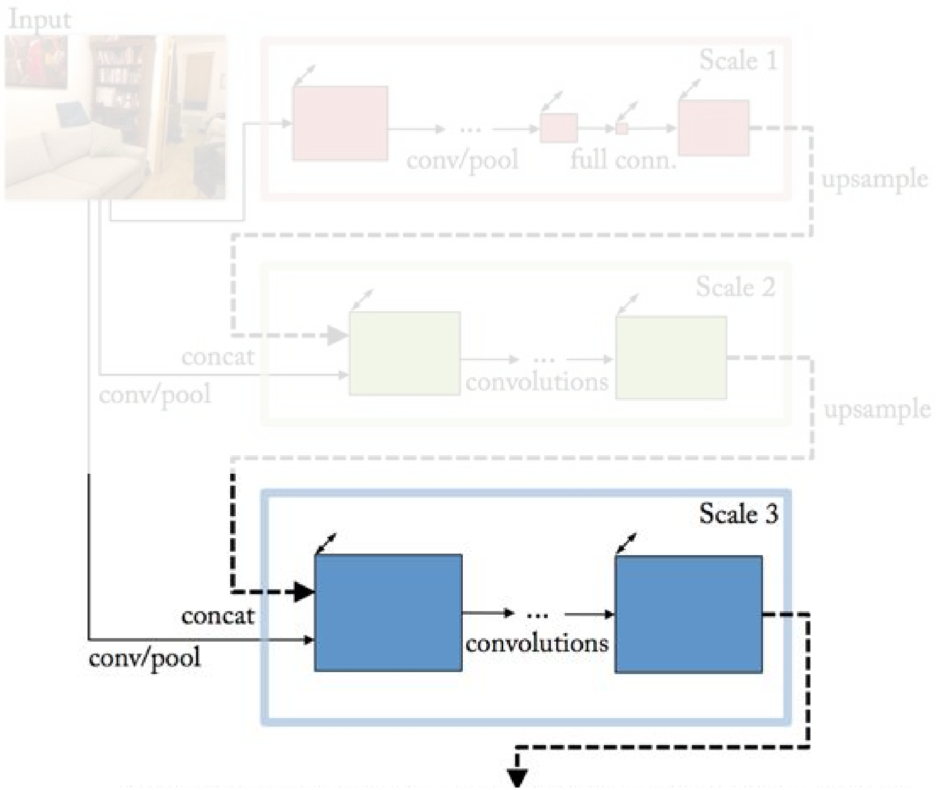
- Coarsest scale
- Full-image field of view at the coarsest scale (1/16 scale)
- AlexNet (smaller model size) and VGG
- Upsampling to be the same size as the input for the next scale
- Important for producing smooth output

Scale 2



- Prediction at scale 2 (¼ scale)
- 5 layers
- Concatenate the output of previous scale with those from a single layer of convolution and pooling
- Output has the same size, with the number of channel depending on the task

Scale 3



- Refines to higher resolution at scale 3 (½ scale)
- 4 layers
- Concatenate the output of scale 2
- Provides details while maintaining the spatially coherent structure from previous scales

Loss Function

- (log) Depth
  Use a scale-invariant error

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

$$L_{depth}(D, D^*) = \frac{1}{n} \sum_i d_i^2 - \frac{1}{2n^2} \left(\sum_i d_i \right)^2 + \frac{1}{n} \sum_i [(\nabla_x d_i)^2 + (\nabla_y d_i)^2]$$

Regularization

- Surface Normal

$$L_{normals}(N, N^*) = -\frac{1}{n} \sum_i N_i \cdot N_i^* = -\frac{1}{n} N \cdot N^*$$

- Semantic Labels
  - C_i: softmax

$$L_{semantic}(C, C^*) = -\frac{1}{n} \sum_i C_i^* \log(C_i)$$

Training

- Use SGD to train scale 1 and 2 jointly; then fix the parameters to train scale 3
- For scale 3, backprop with cropped images (increased stochasticity and efficiency)
- Data augmentation with random linear transformations, colors, contrast
- Parameter sharing in scale 1 for depth and normal networks
- Make use of depth and normal information by applying conv to each input separately

Metrics for evaluation

- Dataset: NYU Depth
- Depth:
  - abs/sqr relative difference
  - RMS (linear, log)
  - Scale invariant difference
- Surface normal: angle distance
- Semantic labelling: pixel-wise, per class accuracy; Jaccard Index

$\delta < 1.25$
$\delta < 1.25^2$
$\delta < 1.25^3$
abs rel
sqr rel
RMS(lin)
RMS(log)
sc-inv.

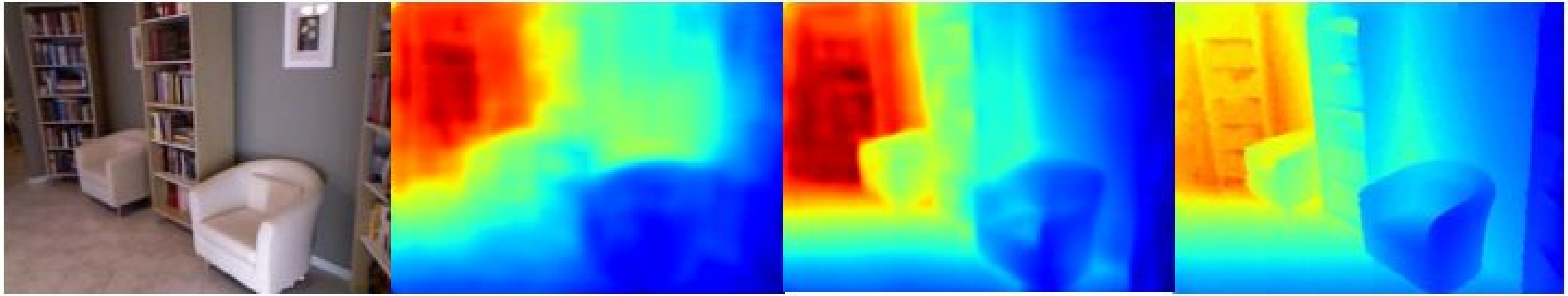
Angle Distance		**Within t° Deg.**		
Mean	**Median**	11.25°	22.5°	30°

Pix. Acc.	Per-Cls Acc.	Freq. Jaccard	Av. Jaccard

Results (Depth)

Outperforms all prior works (Ladicky et al., Karsh et al. [18], Baig et al. [1], Liu et al. [23] and Eigen et al. [8])

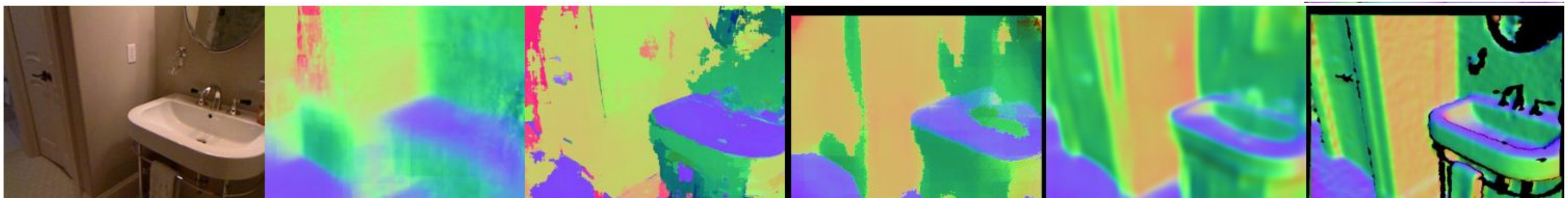
VGG outperforms AlexNet due to larger model size



Results (Surface Normal)

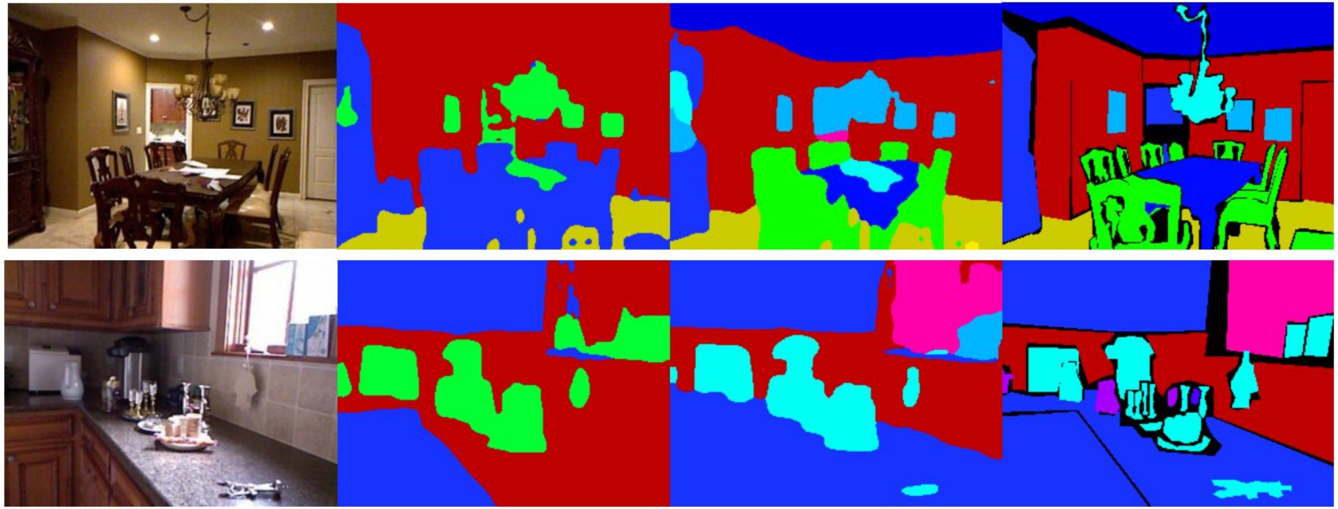
- Results using the AlexNet version for scale 1 are comparable to the prior works (specialized in finding normals)
- Outperforms 3DP, Ladicky et al., Fouhey et al., Wang et al.
- VGG version achieves the best results

Surface Normal Estimation (GT [21])					
	Angle Distance		Within t° Deg.		
	Mean	Median	11.25°	22.5°	30°
3DP [10]	35.3	31.2	16.4	36.6	48.2
Ladicky &*al.* [21]	33.5	23.1	27.5	49.0	58.7
Fouhey &*al.* [11]	35.2	17.9	40.5	54.1	58.9
Wang &*al.* [38]	26.9	14.8	42.0	61.2	68.2
Ours (AlexNet)	23.7	15.5	39.2	62.0	71.1
Ours (VGG)	**20.9**	**13.2**	**44.4**	**67.2**	**75.9**



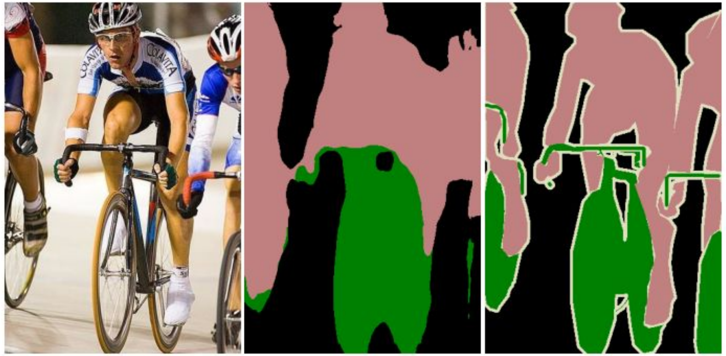
Results (Semantic Labelling)

40-Class Semantic Segmentation				
	Pix. Acc.	Per-Cls Acc.	Freq. Jaccard	Av. Jaccard
Gupta&al.'13 [13]	59.1	28.4	45.6	27.4
Gupta&al.'14 [14]	60.3	35.1	47.0	28.6
Long&al. [24]	65.4	**46.1**	49.5	34.0
Ours (AlexNet)	62.9	41.3	47.6	30.8
Ours (VGG)	**65.6**	45.1	**51.4**	**34.1**



Generalizability

- Applied to the Sift Flow and Pascal VOC dataset
- No need to adjust convolutional kernel size and learning rate



Pascal VOC Semantic Segmentation						
	2011 Validation				2011 Test	2012 Test
	Pix. Acc.	Per-Cls Acc.	Freq.Jacc	Av.Jacc	Av.Jacc	Av.Jacc
Dai&al.[7]	–	–	–	–	–	61.8
Long&al.[24]	90.3	75.9	83.2	62.7	**62.7**	62.2
Chen&al.[5]	–	–	–	–	–	**71.6**
Ours (VGG)	90.3	72.4	82.9	62.2	62.5	62.6

Experiments on architecture

- Scale 1: significant contribution for estimation of all three tasks
- Scale 2: important to semantic labelling; it integrates depth and normal info
- Information provided by scale 2 could be redundant given the predicted D&N

Contributions of Scales

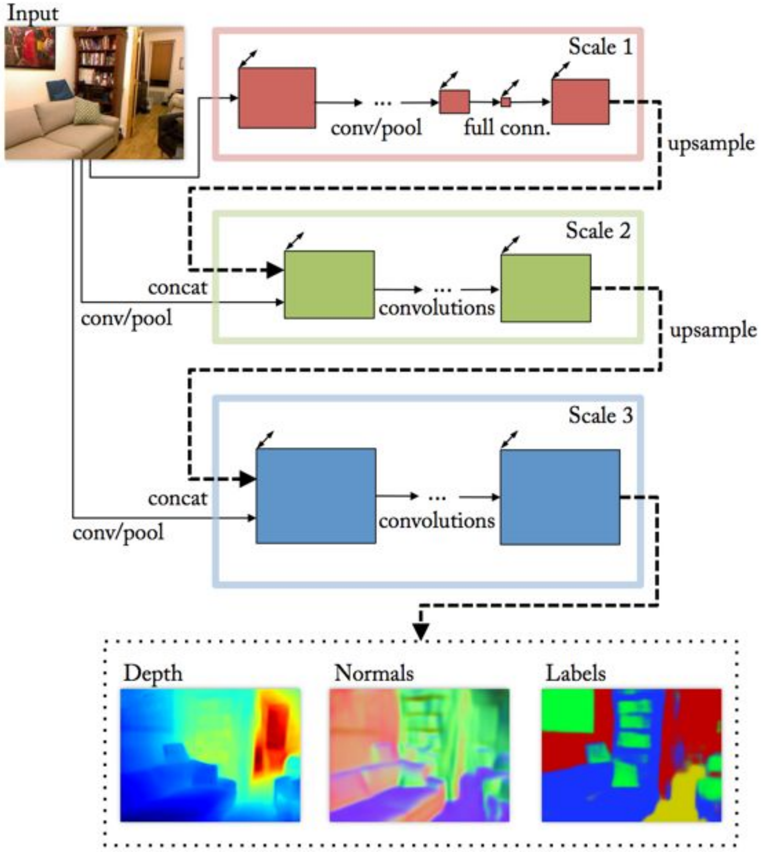
	Depth	Normals	4-Class RGB+D+N	4-Class RGB	13-Class RGB+D+N	13-Class RGB
	Pixelwise Error lower is better		Pixelwise Accuracy higher is better			
Scale 1 only	0.218	29.7	71.5	71.5	58.1	58.1
Scale 2 only	0.290	31.8	77.4	67.2	65.1	53.1
Scales 1 + 2	0.216	26.1	80.1	74.4	69.8	63.2
Scales 1 + 2 + 3	0.198	25.9	80.6	75.3	70.5	64.0

Effect of Depth/Normals Inputs

	Scale 2 only Pix. Acc.	Scale 2 only Per-class	Scales 1 + 2 Pix. Acc.	Scales 1 + 2 Per-class
RGB only	53.1	38.3	63.2	50.6
RGB + pred. D&N	58.7	43.8	65.0	49.5
RGB + g.t. D&N	65.1	52.3	69.8	58.9

Summary

- Depth, surface normals, semantic segmentation from a single image
- Uniform framework
- Multi-scale architecture
- Coarse to fine prediction



Following Works

Semantic Segmentation with ConvNets

Chen, Liang-Chieh, et al. "Semantic image segmentation with deep convolutional nets and fully connected crfs." arXiv preprint arXiv:1412.7062 (2014).

Lin, Guosheng, Chunhua Shen, and Ian Reid. "Efficient piecewise training of deep structured models for semantic segmentation." arXiv preprint arXiv:1504.01013 (2015).

Instance Segmentation with ConvNets

Physical Property Estimation from A Single Image, e.g. intrinsics

Narihira, Takuya, Michael Maire, and Stella X. Yu. "Direct intrinsics: Learning albedo-shading decomposition by convolutional regression." Proceedings of the IEEE International Conference on Computer Vision. 2015.