Factoring Scenes into 3D Structure and Style

David Fouhey

Thesis Committee:

Abhinav Gupta (Co-Chair)

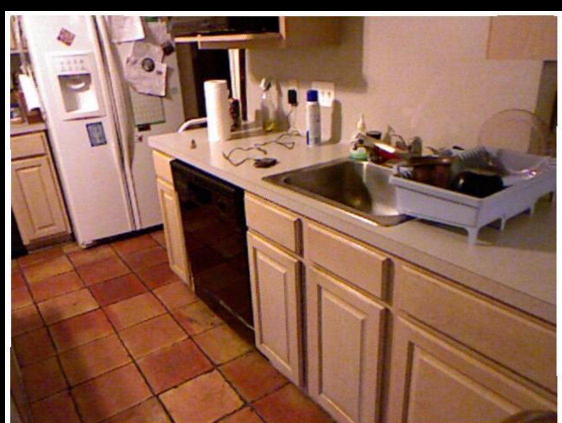
Martial Hebert (Co-Chair)

Deva Ramanan

William T. Freeman, Massachusetts Institute of Technology

Andrew Zisserman, University of Oxford

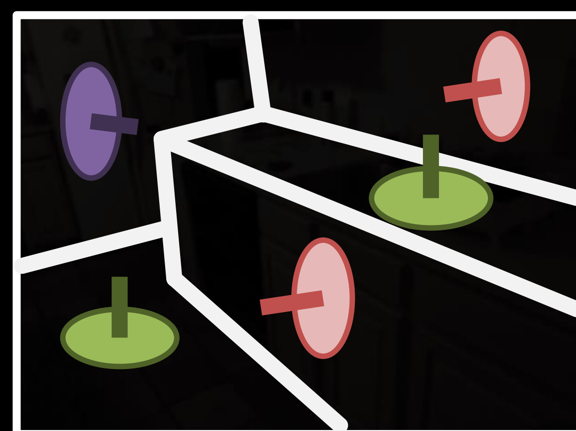
Image



=

3D Structure

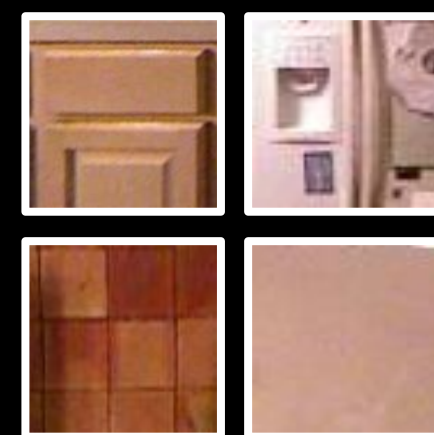
What surfaces are where /
Underlying scene geometry



x

Style

Viewpoint-independent/
canonical texture
(fronto-parallel)

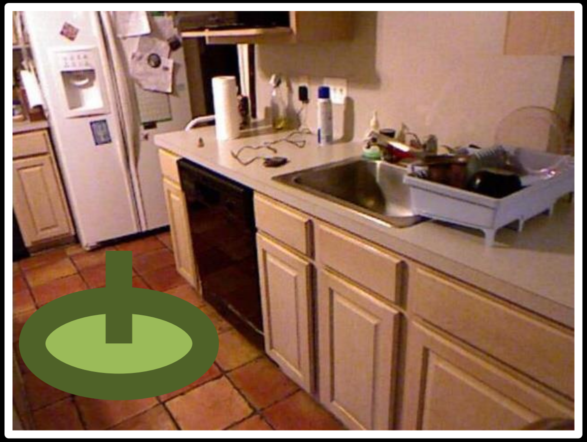


Example

Image

=

3D Structure



x

Style

You See…

Unfortunately…

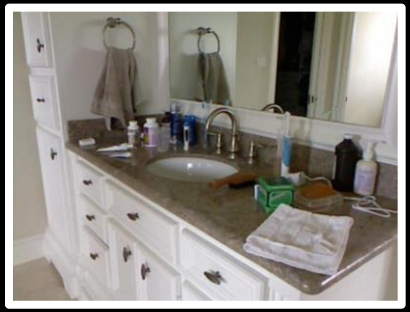
Why Can We Solve It?

Not all factorizations are equally likely!

3D Structure

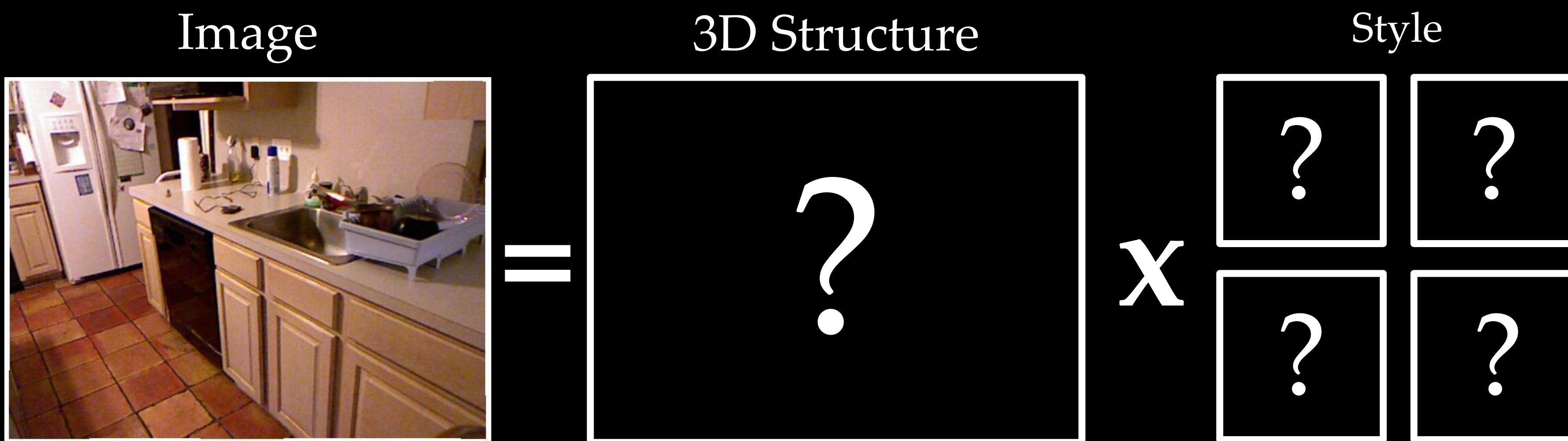
Style

Why Can We Solve It?



…

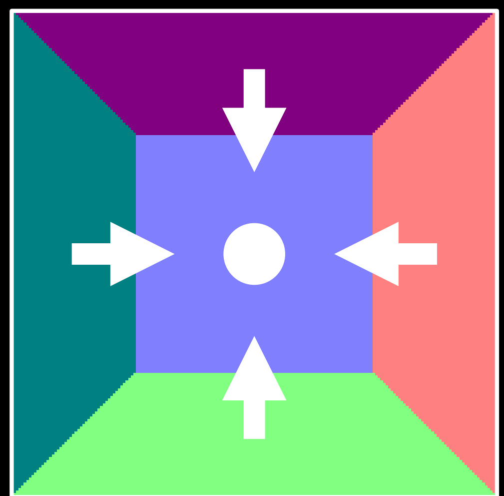
The Problem



Representations/Visualization

3D Structure

Sample
Room

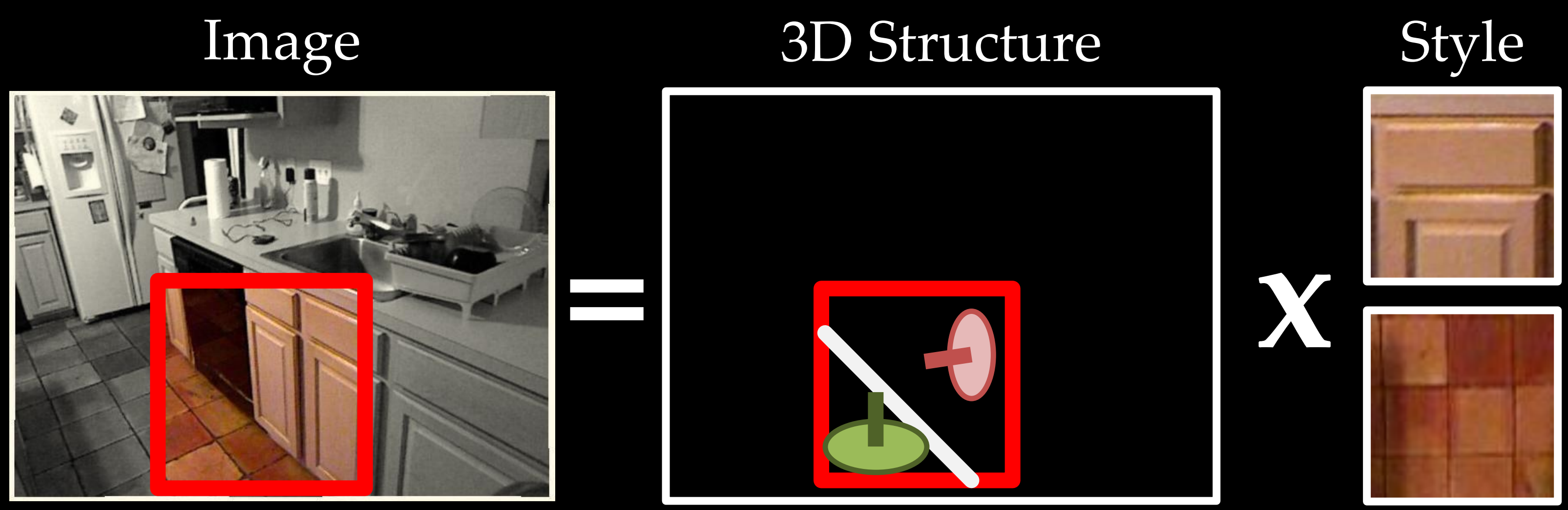


Surface Normal
Legend

Style

Contributions

Our First Contribution

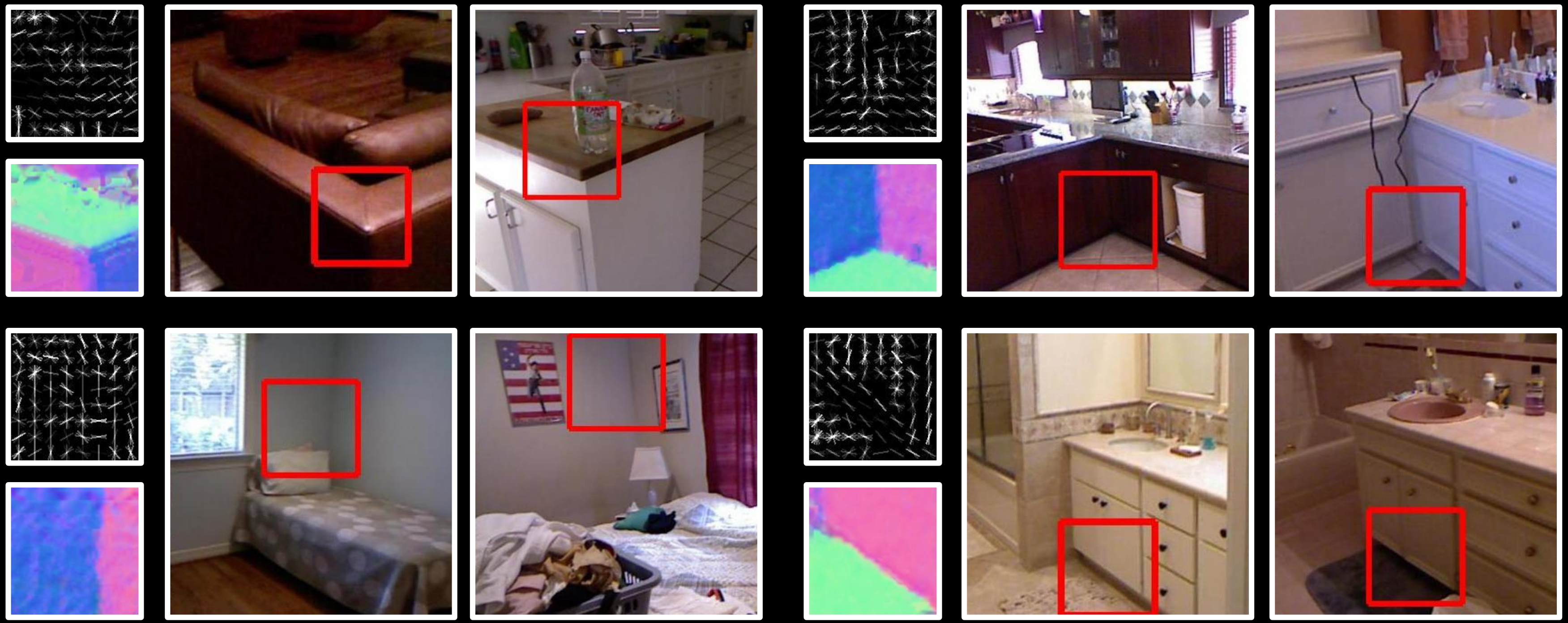


Data-Driven 3D Primitives for Single Image Understanding.
Fouhey, Gupta, Hebert. In ICCV '13.

Supervised Approach

Data-Driven 3D Primitives for Single Image Understanding.
Fouhey, Gupta, Hebert. In ICCV '13.

Supervised Approach



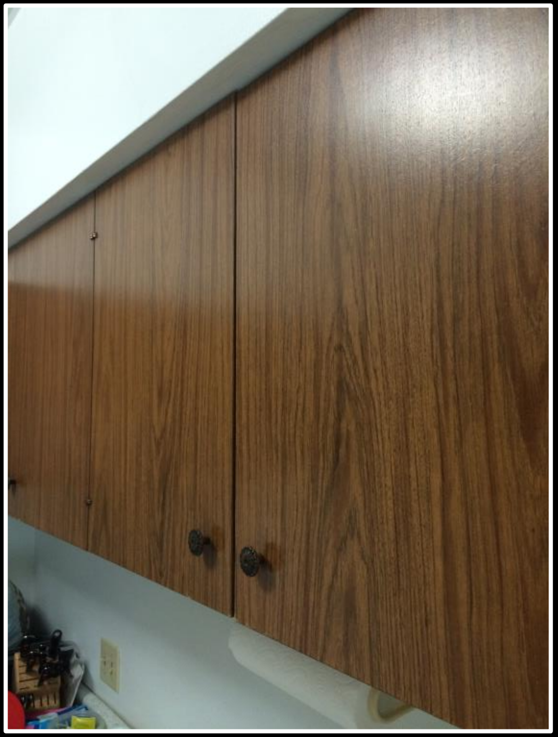
Data-Driven 3D Primitives for Single Image Understanding.
Fouhey, Gupta, Hebert. In ICCV '13.

Supervised Approach

Data-Driven 3D Primitives for Single Image Understanding.
Fouhey, Gupta, Hebert. In ICCV '13.

Issue #1 – Data

Wasteful: no cross-viewpoint sharing



Solution

Explicit factorization via style elements:
cross-viewpoint and *do not require training data*

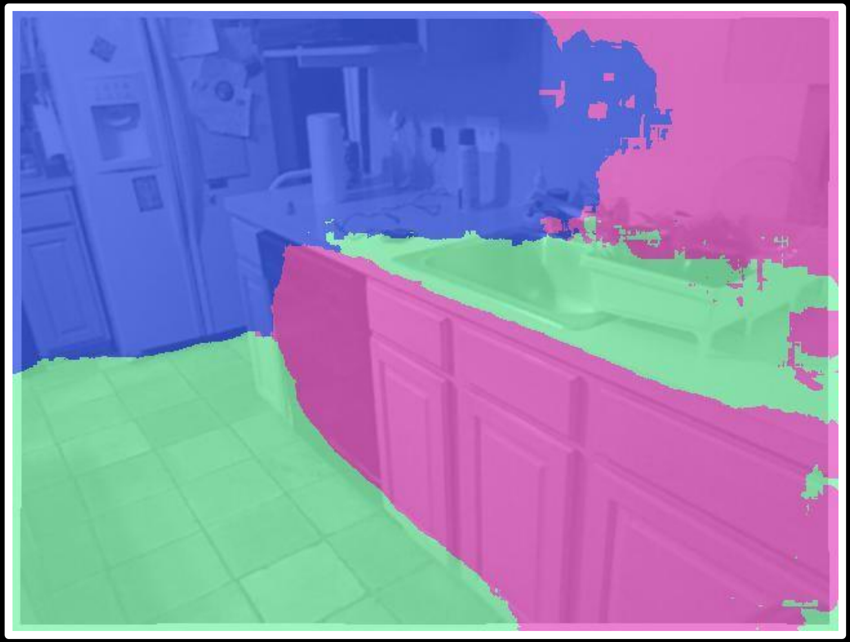
Style Element

Detections

Single Image 3D Without a Single 3D Image.
Fouhey, Hussain, Gupta, Hebert. In ICCV '15.

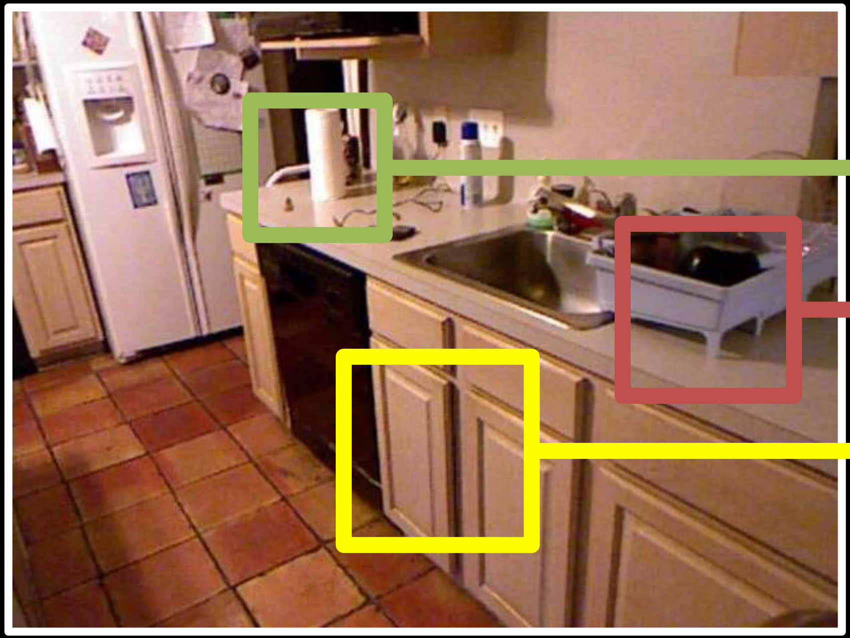
Issue #2

When do we apply domain knowledge/constraints?



Solution

Higher-order Shape Properties



“Cylindrical”

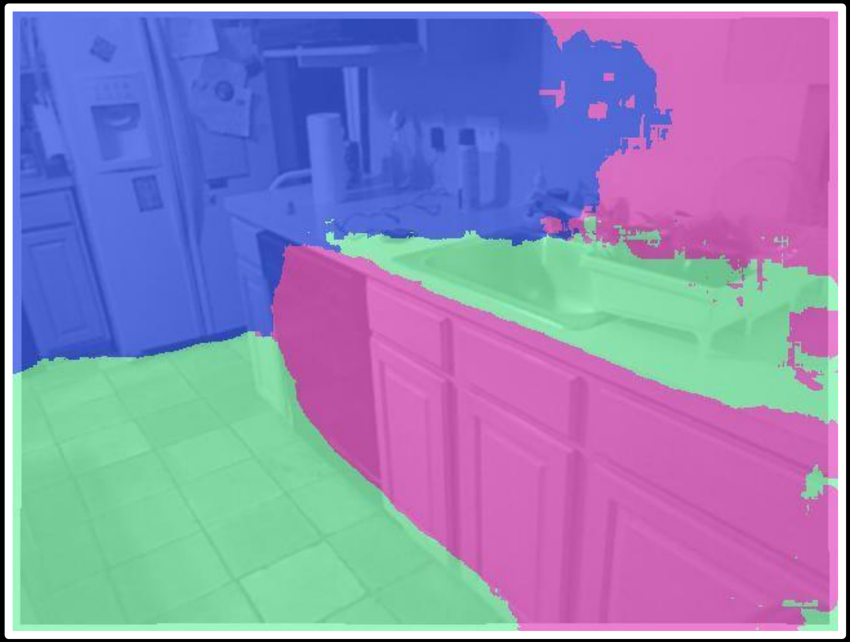
“Point Contact”

“Planar”

3D Shape Attributes.
Fouhey, Gupta, Zisserman. In CVPR ’16.

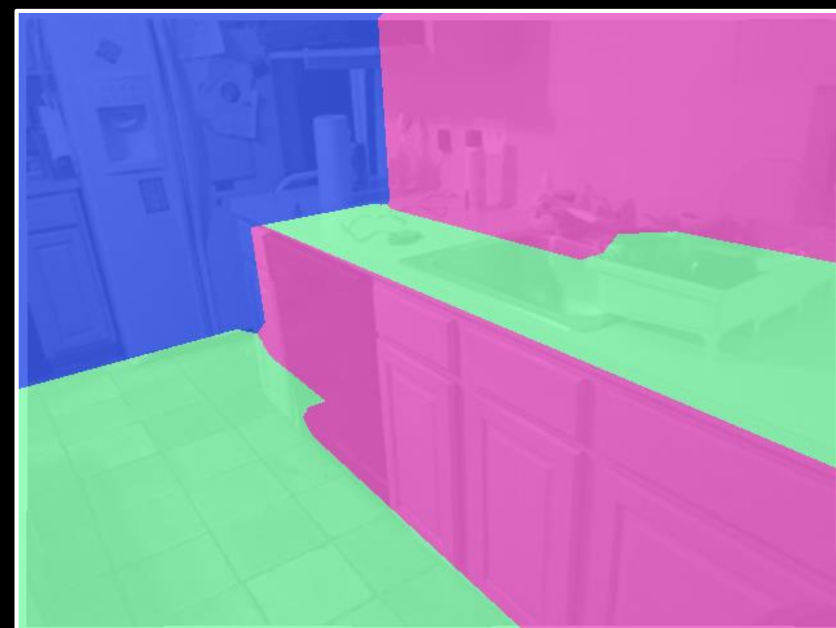
Issue #3

World is much more constrained than per-pixel but more detailed than global properties.



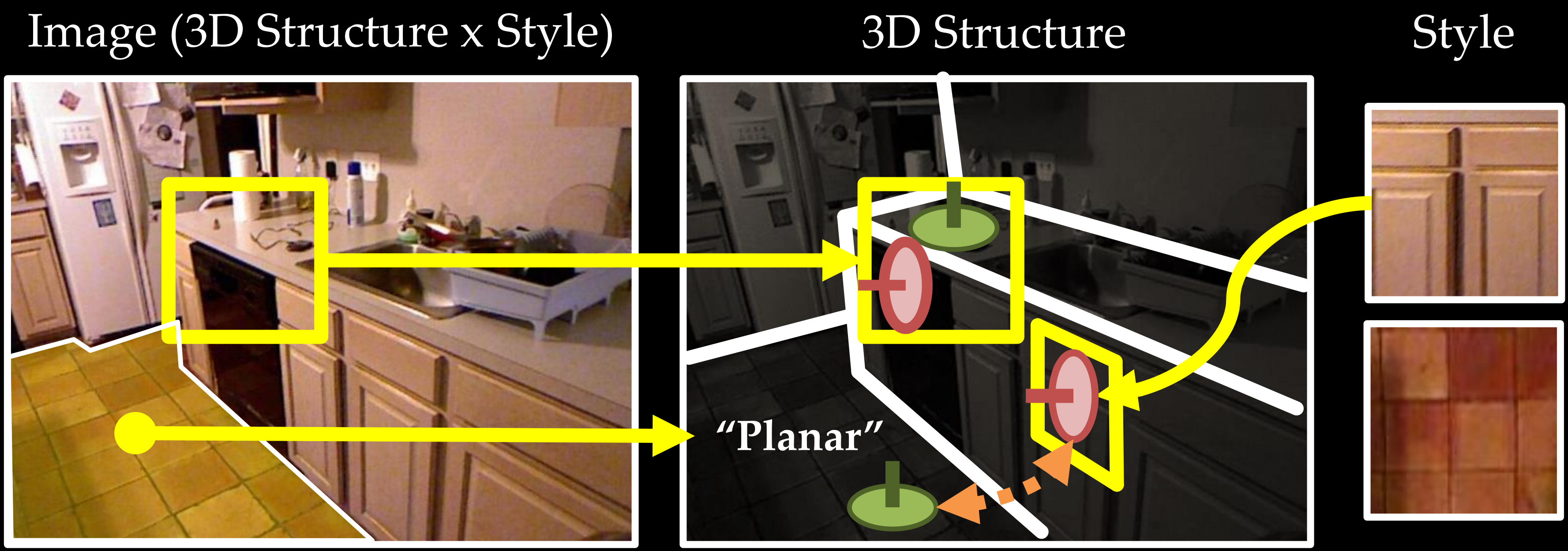
Solution

Mid-level constraints, discrete scene parses



Unfolding an Indoor Origami World.
Fouhey, Gupta, Hebert. In ECCV '14.

Dissertation Contributions

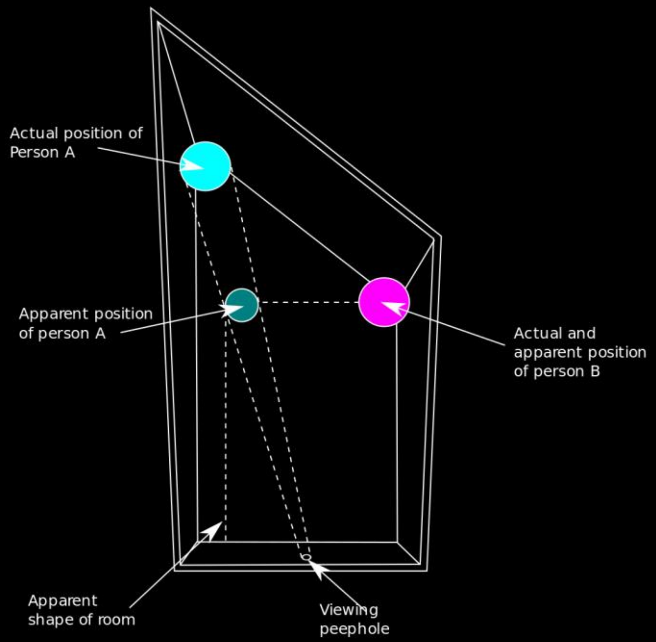


1. Local image-based cues
2. Local style-based cues
3. Cues for higher-order 3D structure
4. Constraints on 3D structure
5. Data-driven dense normal estimation as a scene understanding task

RELATED WORK

Human Vision

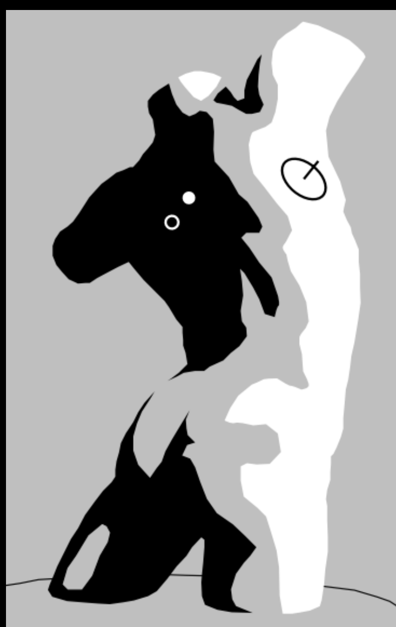
- Monocular cues are integral to “normal” vision
- Monocular can override binocular: monocular illusions persist under binocular conditions

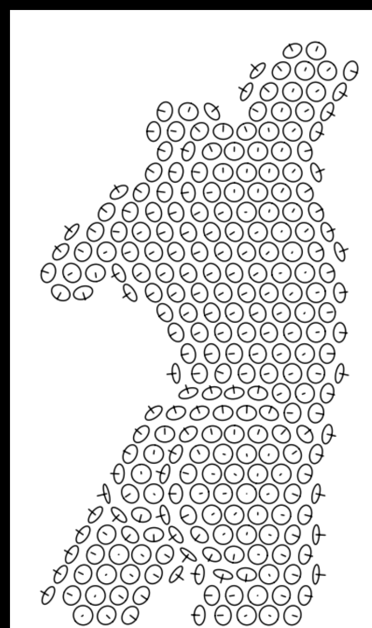


Gehringer and Engel, Journal of Experimental Psychology: Human Perception and Performance, 1986

Human Vision

Higher order properties are ***not*** obtained from depthmaps





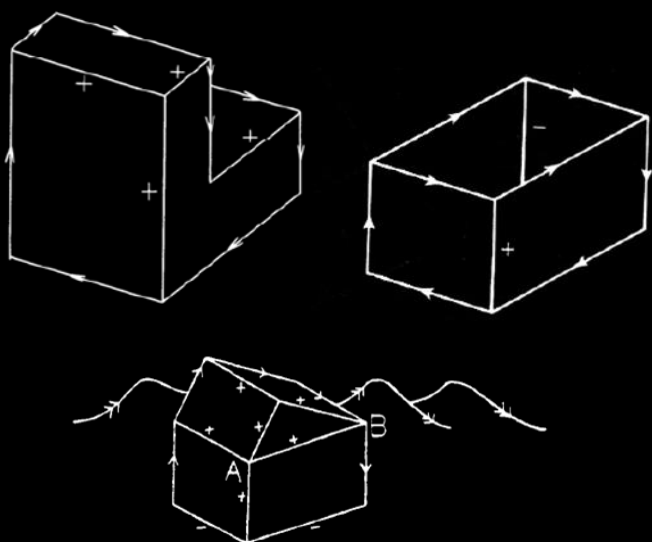
“It is rather unlikely that the attitudes [i.e.,normals] are derived from a pictorial depthmap”
-Koenderink, van Doorn, Kappers ‘96

“Judgements about the curvature of local surface patches were too precise to be based on a symbolic representation of surface orientation ”
-Johnston and Passamore, ‘93

Koenderink, van Doorn, Kappers, Pictorial surface attitude and local depth comparisons. *Perception and Psychophysics*, 1996
Johnston and Passamore, . Independent encoding of surface orientation and surface curvature, *Vision Research*, 1994

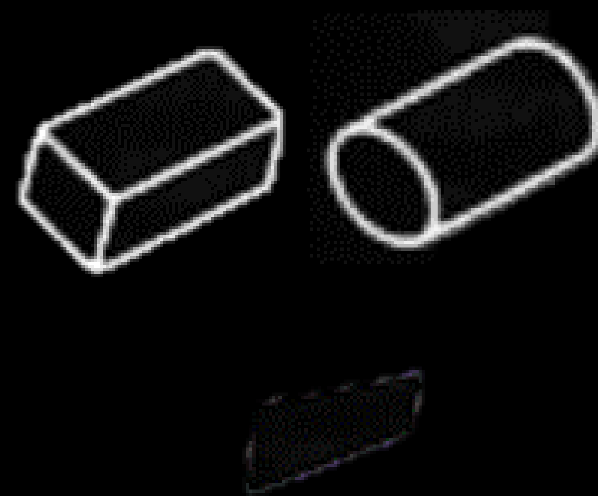
Recovering 3D Structure

Line-Based Primitives



Roberts 1963, Guzman 1968, Huffman 1971, Clowes 1971, Waltz, 1975, Kanade 1980, Sugihara 1986, Malik 1987, etc.

Volumetric Primitives



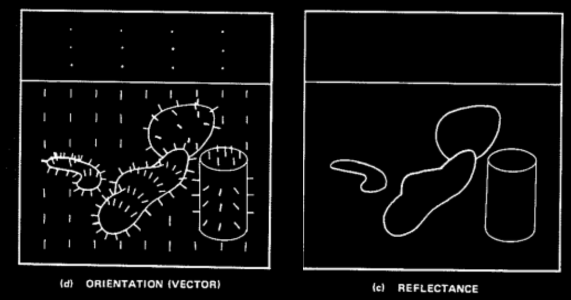
Binford 1971, Brooks 1979, Biederman 1987, etc.

Recovering 3D Structure

Hoiem et al., 2005
Qualitative Orientation

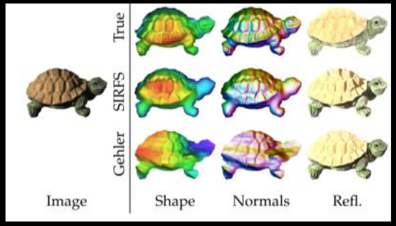
Saxena et al., 2005
Quantitative Depth

Image Factorization



Barrow and Tenenbaum 1978

Shape-from-X



Tappen et al., 2002, 2006, Grosse et al. 2009, Barron et al. 2012, etc.

Malik et al. 1997, Criminisi et al., 2000, Forsyth 2002, Zhang et al. 2014, etc.

Content & Style

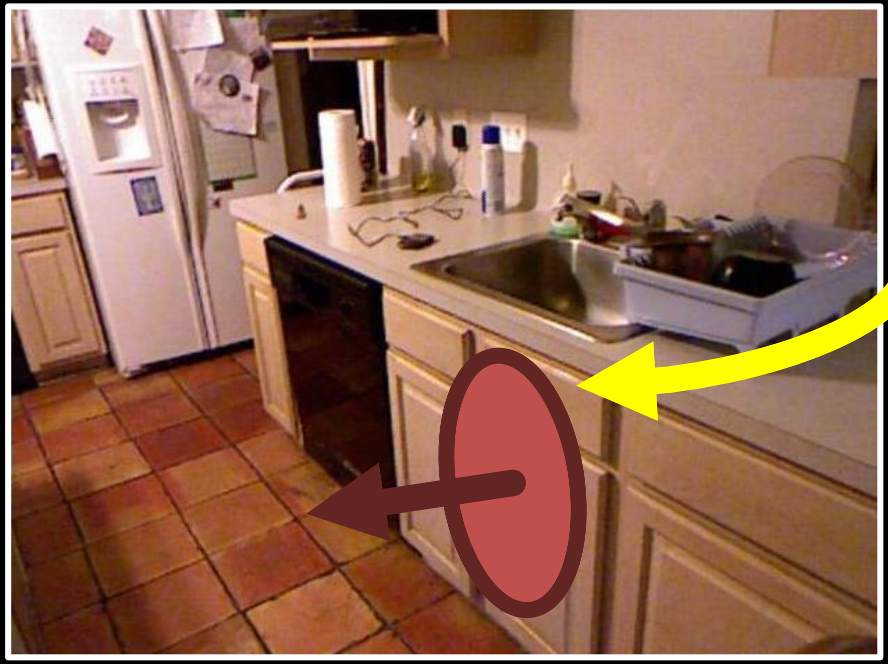
Tenenbaum et al., 1997

Elgammal et al. 2004, Wang et al., 2007, Pirsiavash 2009, etc.

SURFACE NORMALS

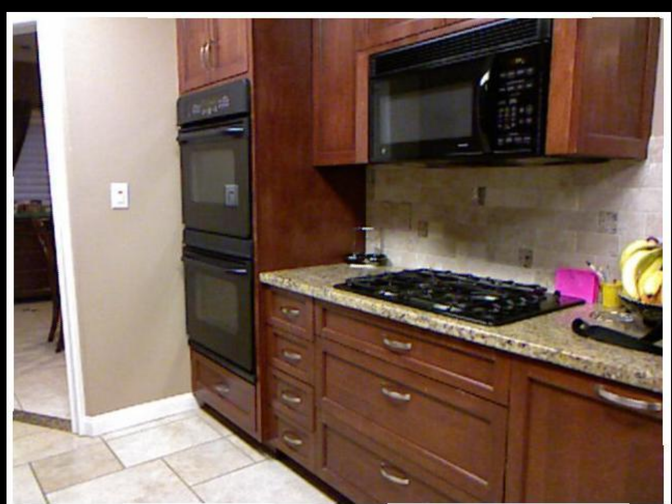
Surface Normals

Quantitative Orientation

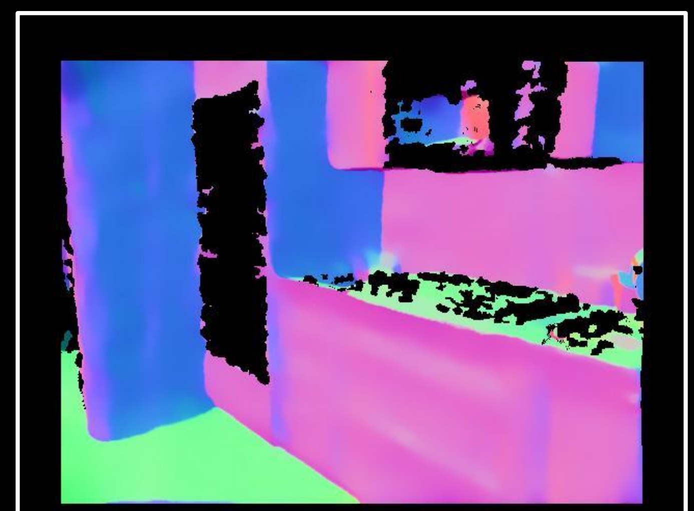


$$\begin{bmatrix} 0.82 \\ -0.21 \\ 0.53 \end{bmatrix}$$

Obtaining Normals



Color Image



Normals

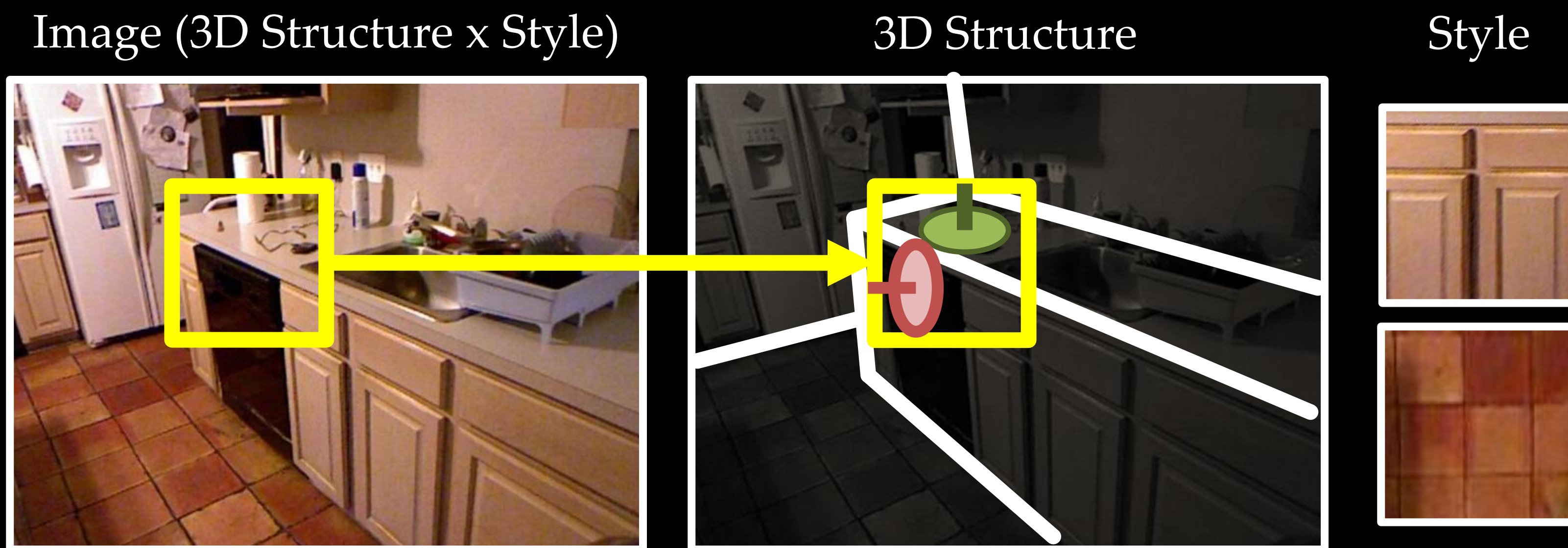
Evaluating Normals

Input GT Prediction

Aggregate over the entire dataset, compute:
mean(**E**), median(**E**), sqrt(mean(**E**)),
mean(**E** < t), t = 11.25, 22.5, 30

Why Normals?

- Direct modeling produces better results
- Observable from perspective cues as opposed to scaling
- Fewer ambiguities than depth



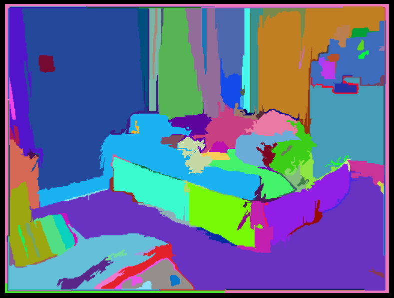
Local image-based cues

DATA-DRIVEN 3D PRIMITIVES

Previous Primitives

Lines + Planes

Kanade 1981,
Sugihara 1986,
Liebowitz et al. 1998,
Criminisi et al. 1999,
Lee et al., 2009, etc.



Segments

Hoiem et al. 2005,
Saxena et al. 2005,
Ramalingam et al. 2008, etc.

Rooms

Hedau et al. 2009,
Flint et al. 2010,
Flint et al. 2011,
Satkin et al. 2012,
Schwing et al. 2012,
etc.

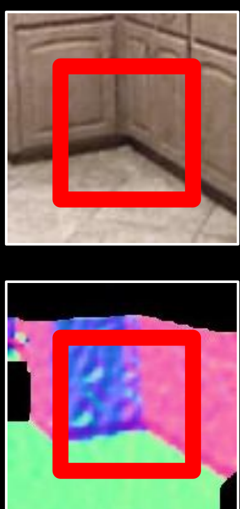
Cuboids

Lee et al. 2010,
Gupta et al. 2010,
Gupta et al. 2011,
Xiao et al. 2012,
Schwing et al. 2013
etc.

Objective

Visually Discriminative

Geometrically Informative

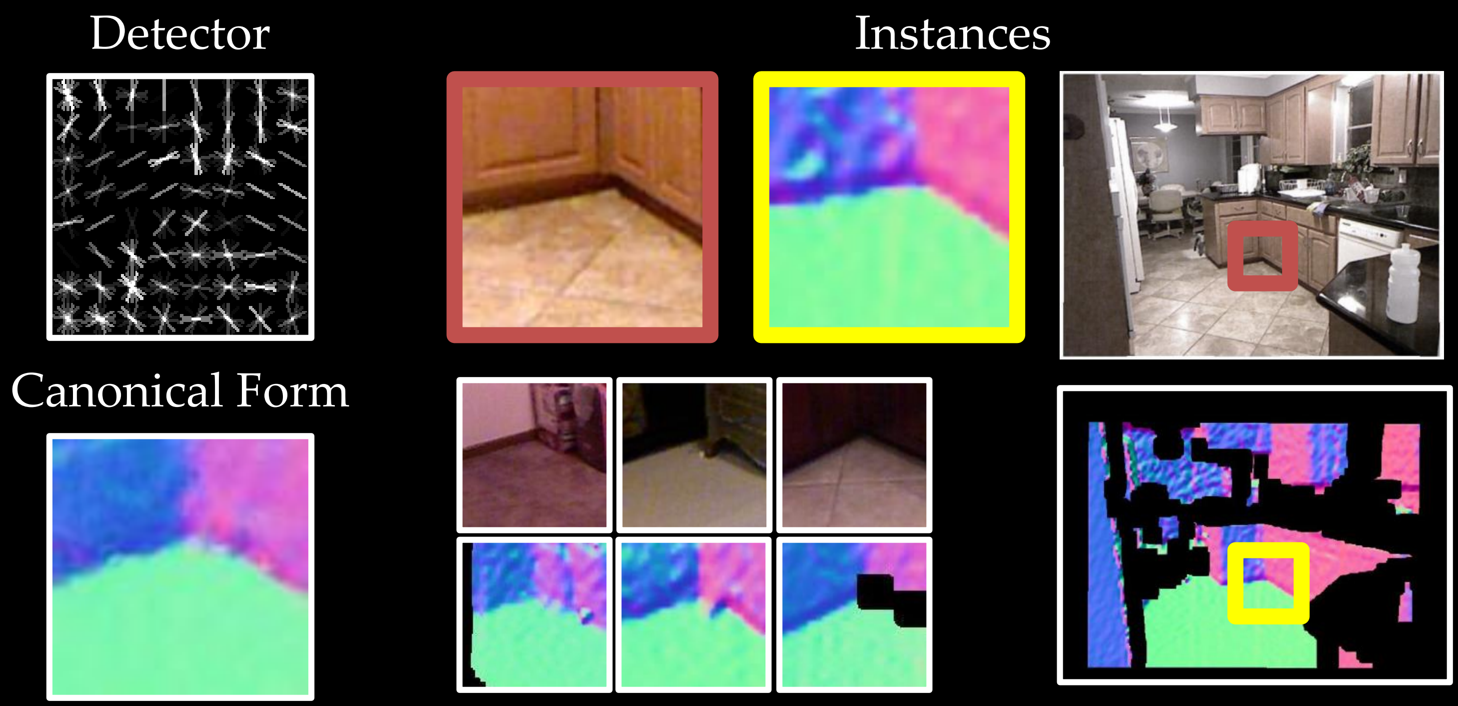


Image

Surface Normals

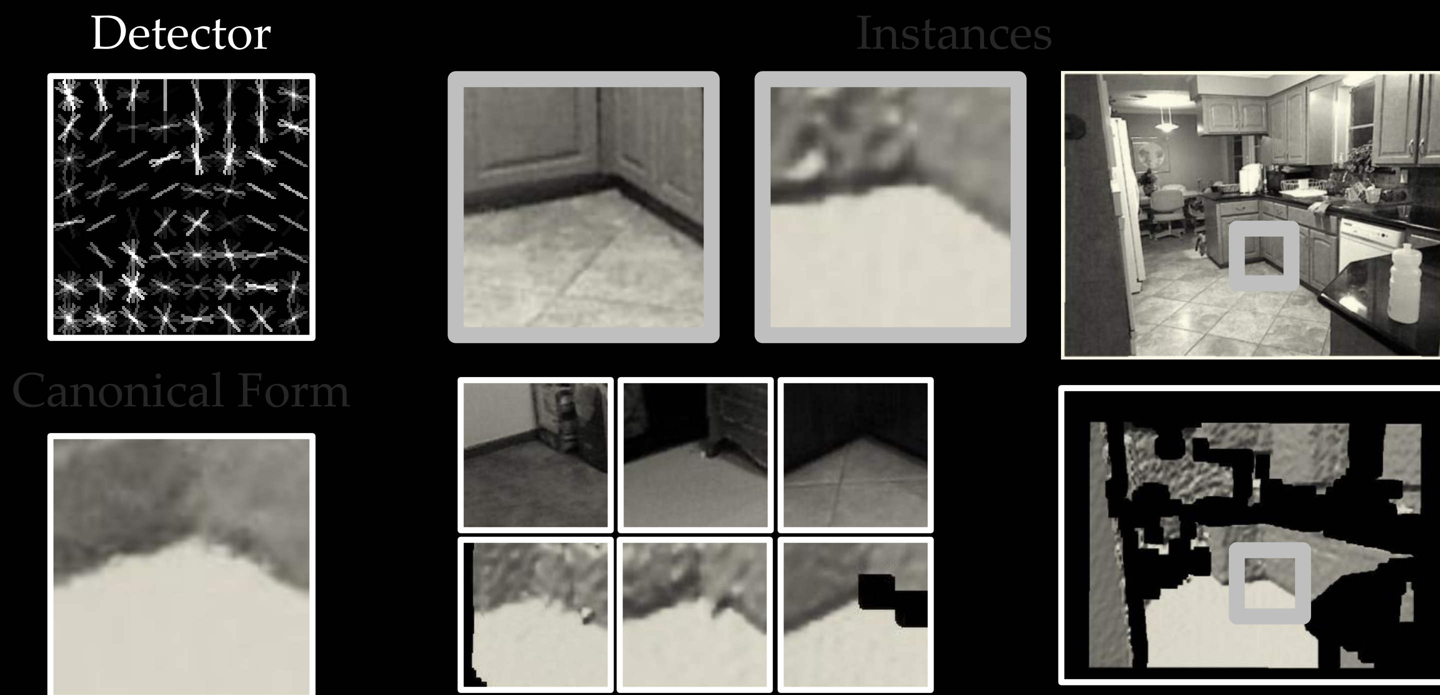
Similar ideas presented concurrently at ICCV '13:
Owens et al., Shape Anchors for Data-Driven Multi-view Reconstruction
Dollar et al., Structured Forests for Fast Edge Detection ;

Representation



Representation

w



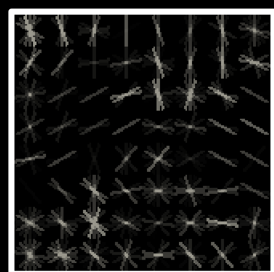
Representation

N

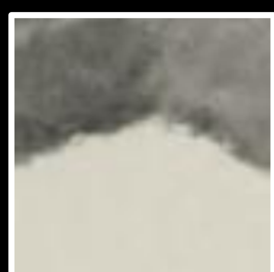
Representation

$\mathbf{y}$

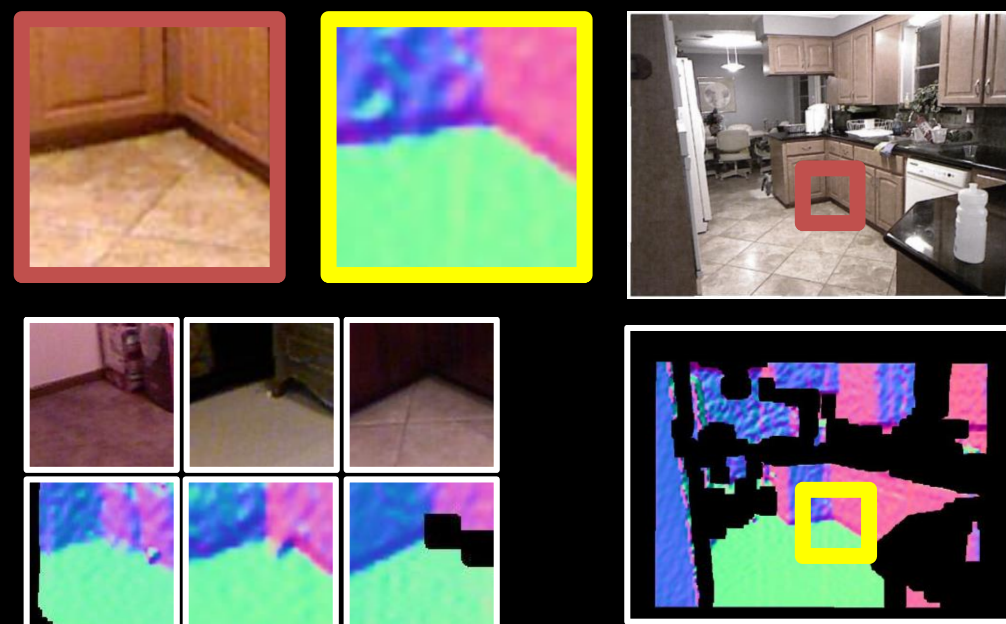
Detector



Canonical Form



Instances



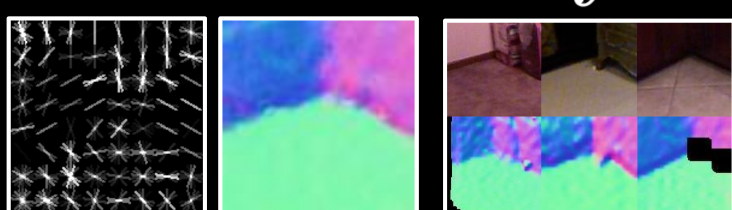
Objective

$$\min_{\mathbf{y},\mathbf{w},\mathbf{N}} R(\mathbf{w}) + \sum_{i=1}^{m} \left[c_2 L(\mathbf{w}, \mathbf{N}, \mathbf{x}_i^A, y_i) + c_1 y_i \Delta(\mathbf{N}, \mathbf{x}_i^G) \right]$$

$$\text{s.t.} |\mathbf{y}|_1 \geq c$$

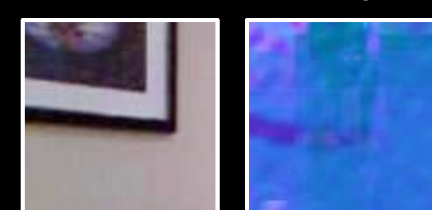
Primitive

$\mathbf{w}$ $\mathbf{N}$ $\mathbf{y}$



Patch

$\mathbf{x}_i^A$ $\mathbf{x}_i^G$



Objective

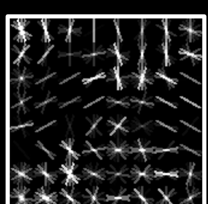
Regularized classifier; loss for
labels determined by geometry

$$\min_{\mathbf{y},\mathbf{w},\mathbf{N}} R(\mathbf{w}) + \sum_{i=1}^{m} \left[c_2 L(\mathbf{w}, \mathbf{N}, \mathbf{x}_i^A, y_i) + c_1 y_i \Delta(\mathbf{N}, \mathbf{x}_i^G) \right]$$

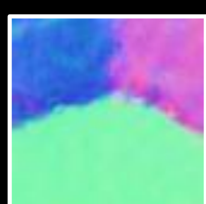
$$\text{s.t.} |\mathbf{y}|_1 \geq c$$

Primitive

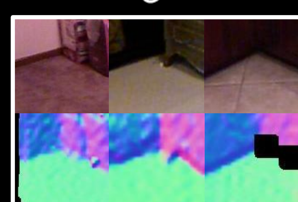
$\mathbf{w}$



$\mathbf{N}$



$\mathbf{y}$



Patch

$\mathbf{x}_i^A$

$\mathbf{x}_i^G$

Objective

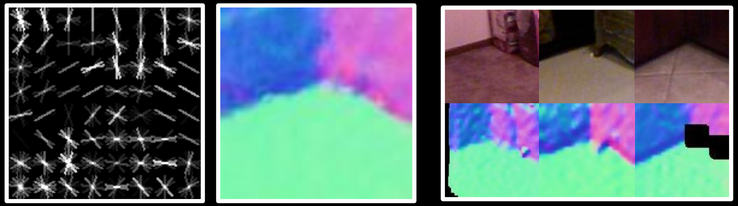
Minimize intra-cluster
geometric distance

$$\min_{\mathbf{y},\mathbf{w},\mathbf{N}} R(\mathbf{w}) + \sum_{i=1}^{m} \left[c_2 L(\mathbf{w}, \mathbf{N}, \mathbf{x}_i^A, y_i) + c_1 y_i \Delta(\mathbf{N}, \mathbf{x}_i^G) \right]$$

$$\text{s.t.} |\mathbf{y}|_1 \geq c$$

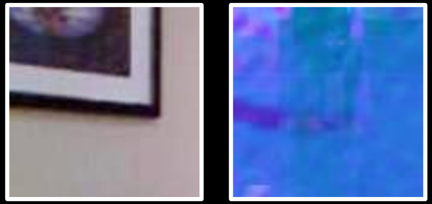
Primitive

$\mathbf{w}$ $\mathbf{N}$ $\mathbf{y}$



Patch

$\mathbf{x}_i^A$ $\mathbf{x}_i^G$



Objective

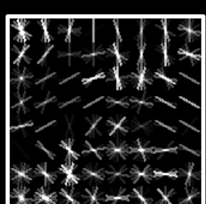
Solve with an approach similar to block-coordinate descent

$$\min_{\mathbf{y},\mathbf{w},\mathbf{N}} R(\mathbf{w}) + \sum_{i=1}^{m} \Big[c_2 L(\mathbf{w}, \mathbf{N}, \mathbf{x}_i^A, y_i) + c_1 y_i \Delta(\mathbf{N}, \mathbf{x}_i^G) \Big]$$

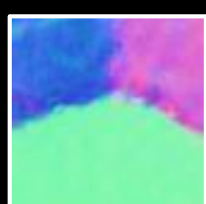
$$\text{s.t.} |\mathbf{y}|_1 \geq c$$

Primitive

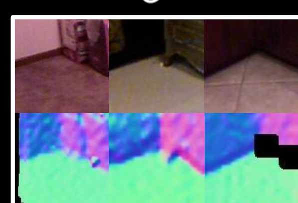
$\mathbf{w}$



$\mathbf{N}$

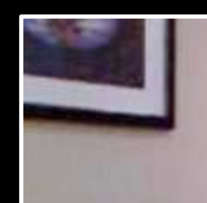


$\mathbf{y}$



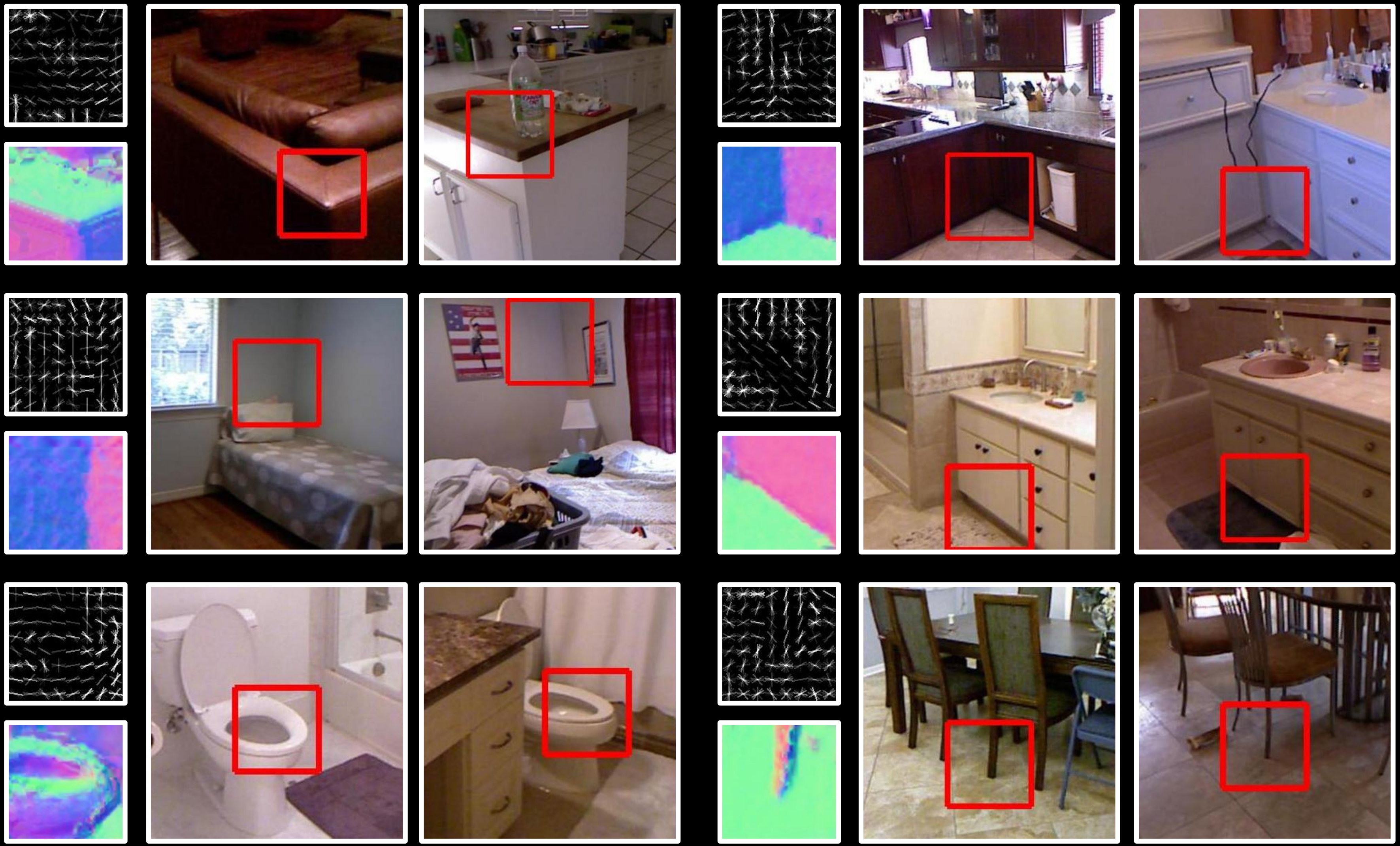
Patch

$\mathbf{x}_i^A$



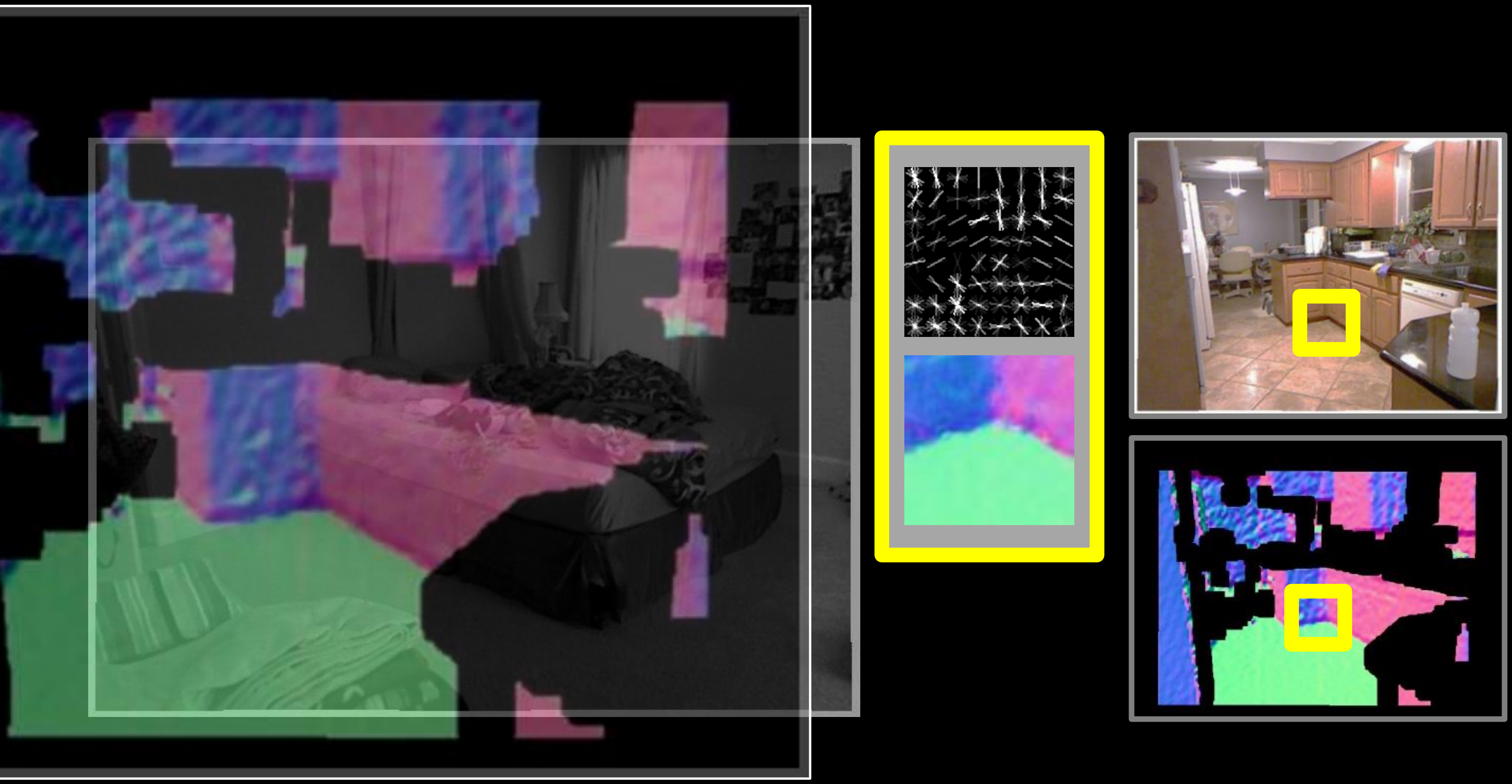
$\mathbf{x}_i^G$

Learned Primitives



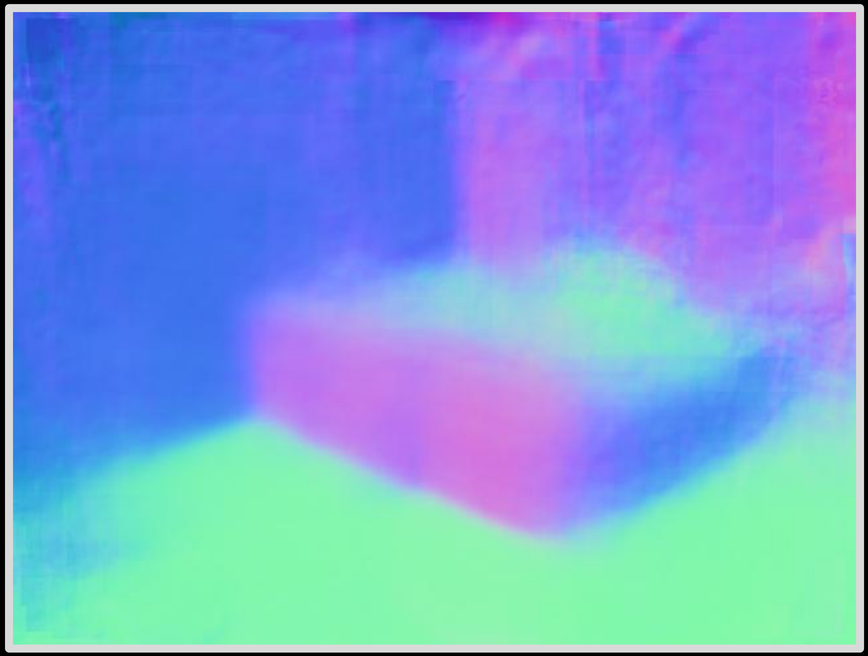
Interpretation from Primitives

Interpretation from Primitives

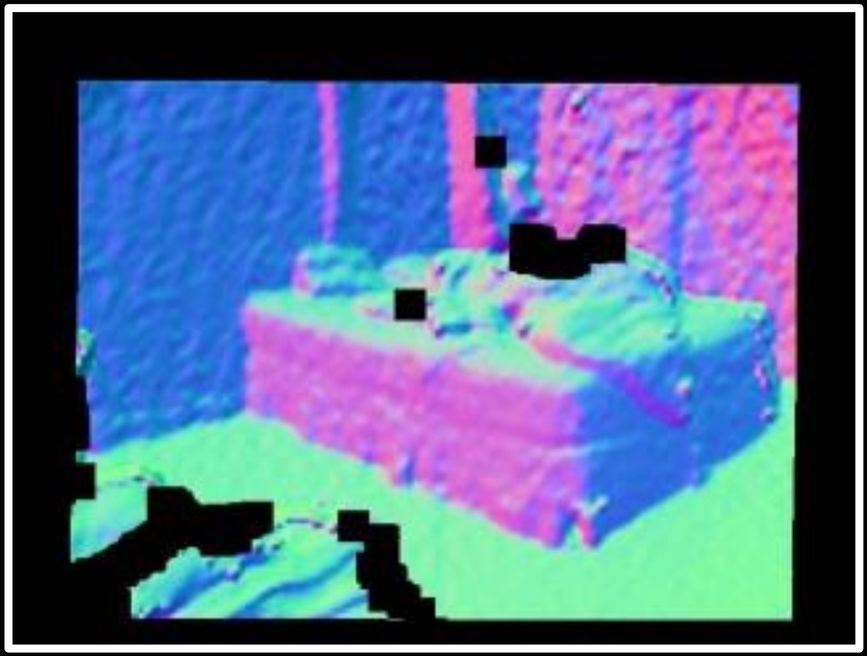


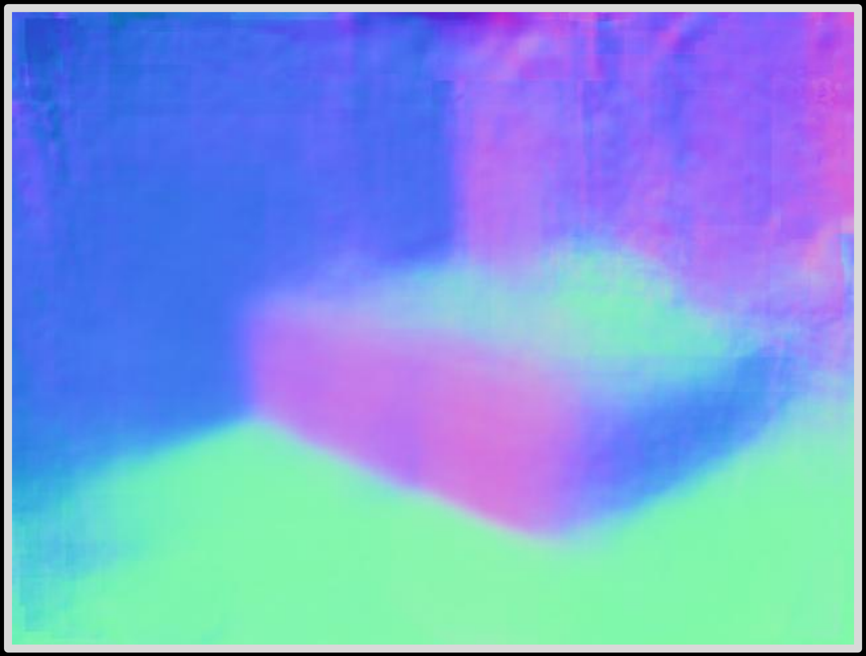
Interpretation from Primitives

Interpretation from Primitives

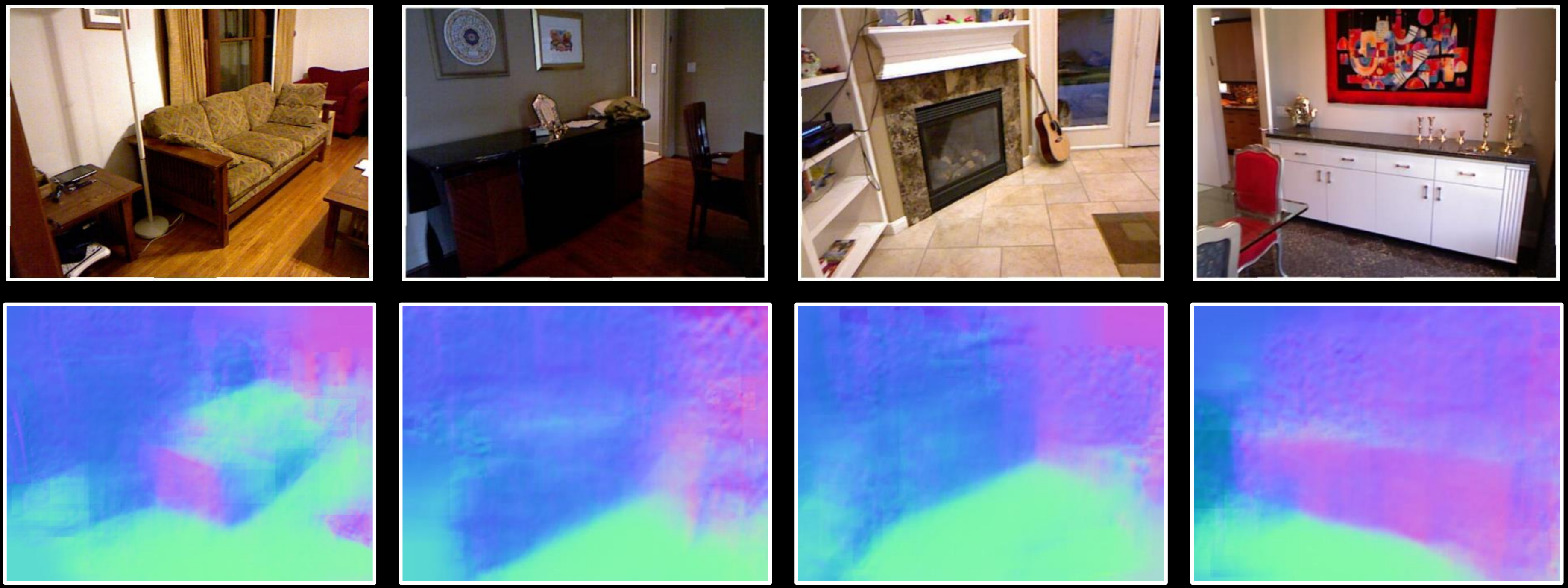


Interpretation from Primitives

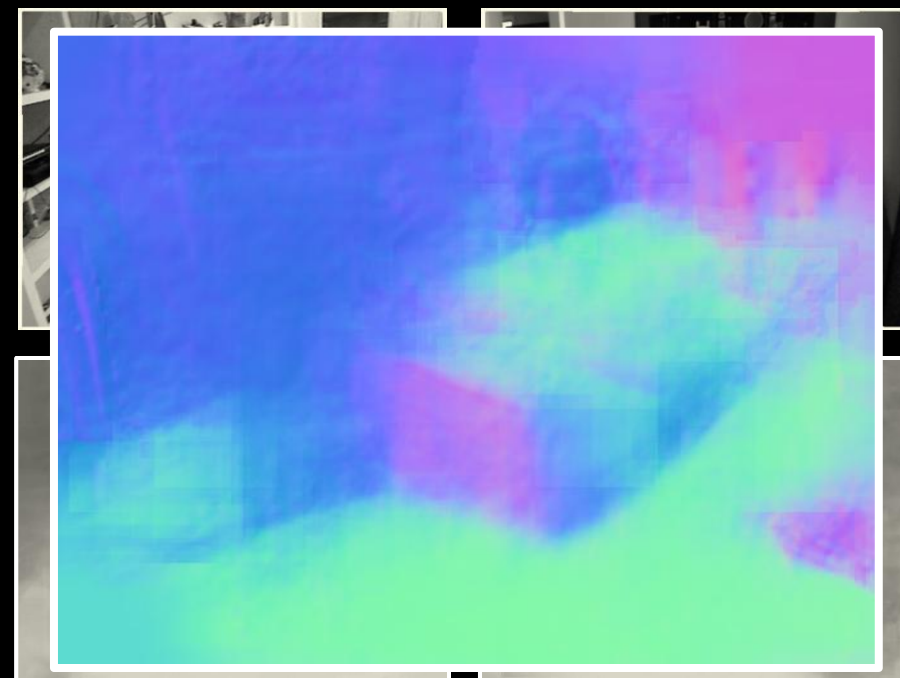




Interpretation from Primitives



Interpretation from Primitives



Results – Quantitative

	Summary Stats (°) (Lower Better)			% Good Pixels (Higher Better)		
	Mean	Median	RMSE	11.25°	22.5°	30°
3DP	**<u>34.2</u>**	**<u>30.0</u>**	**<u>41.4</u>**	**<u>18.6</u>**	**<u>38.6</u>**	**<u>49.9</u>**
Karsch et al.	40.7	37.8	46.9	8.1	25.9	38.2
Saxena et al.	48.0	43.1	57.0	10.7	27.0	36.3
Hoiem et al.	41.2	35.1	49.2	9.0	31.2	43.5
RF+SIFT	36.0	33.4	41.7	11.4	31.4	44.5

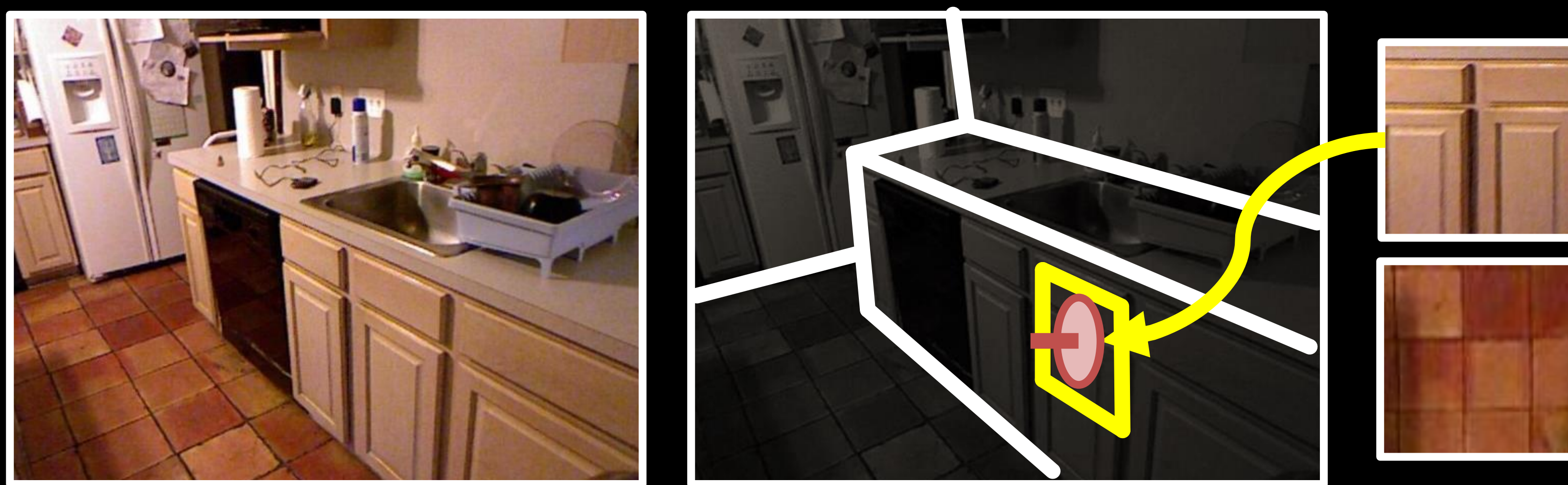
Karsch et al., ECCV 2012; Hoiem et al., ICCV 2005; Saxena et al. NIPS 2005
Fouhey, Gupta, Hebert, ICCV '13.

Issues

Pure memorization: no sharing between views

Learning requires a specialized sensor

Image (3D Structure x Style) 3D Structure Style



Local style-based cues

STYLE ELEMENTS

A Different Idea

These are easy
to get in bulk

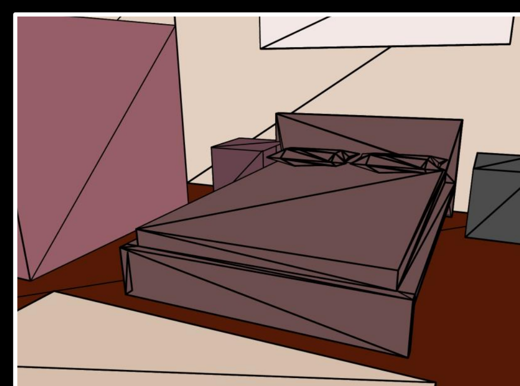
We can put
priors on this

Image

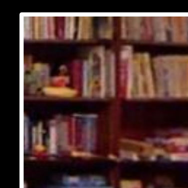
3D Structure

Style

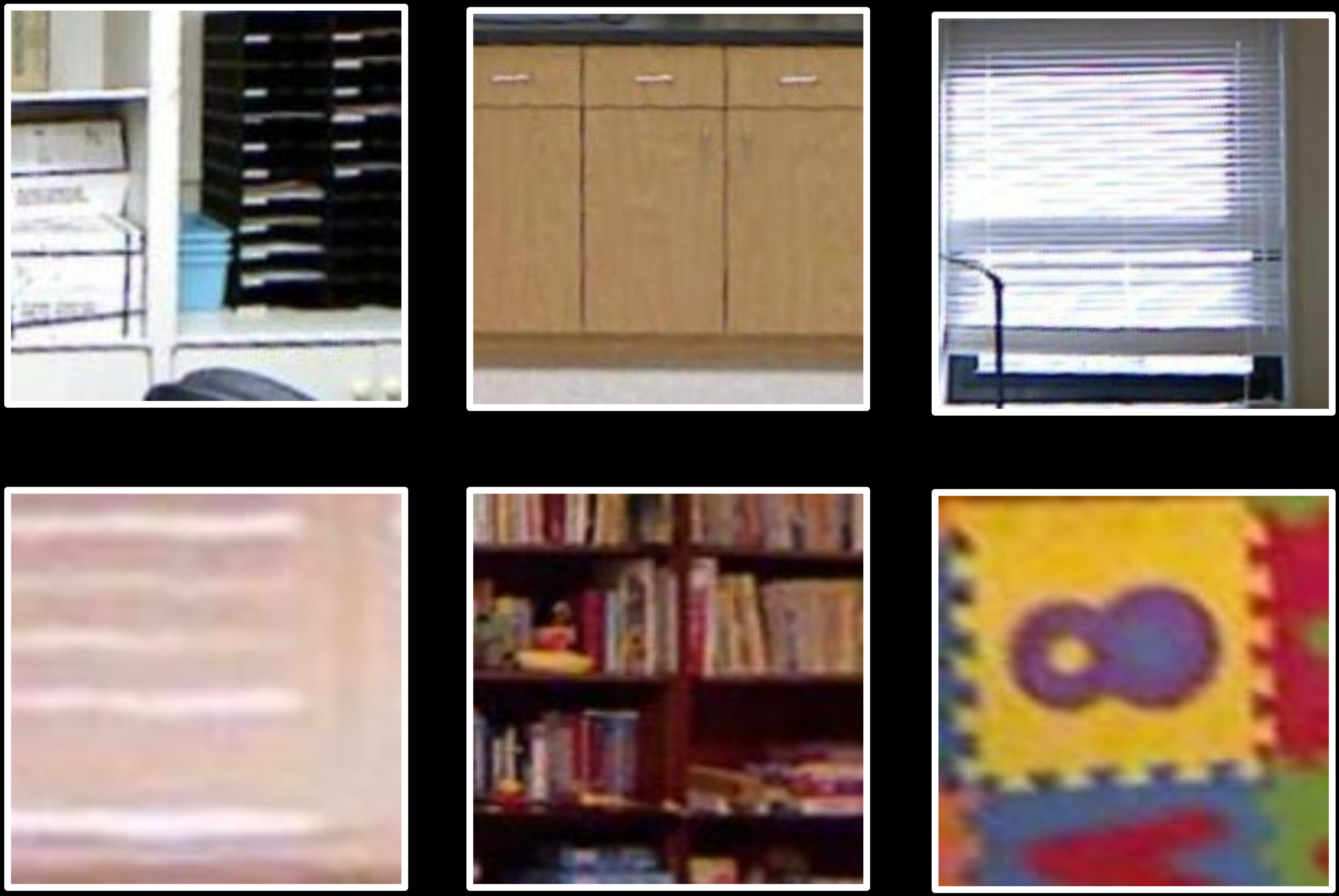
=



x



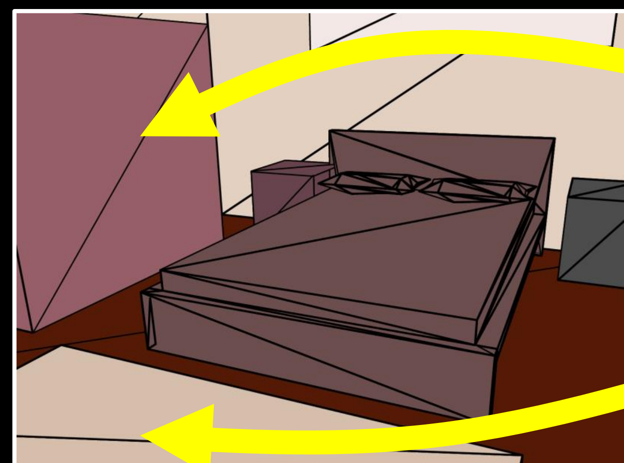
Style Elements



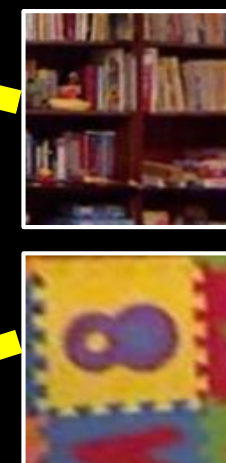
Factorization

Image | 3D Structure | Style

=



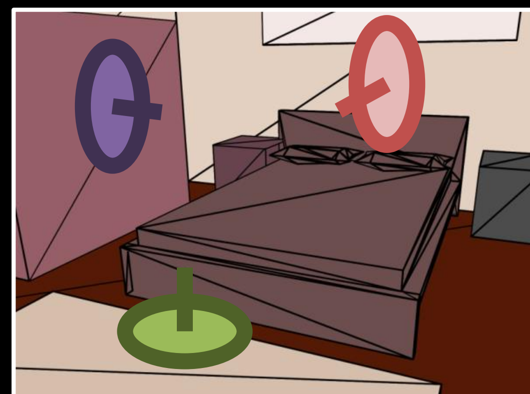
x



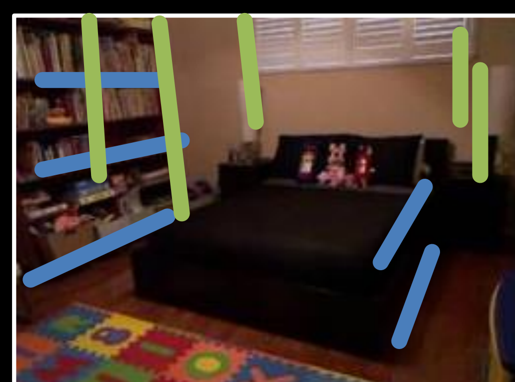
Solving for Style

Image

3D Structure

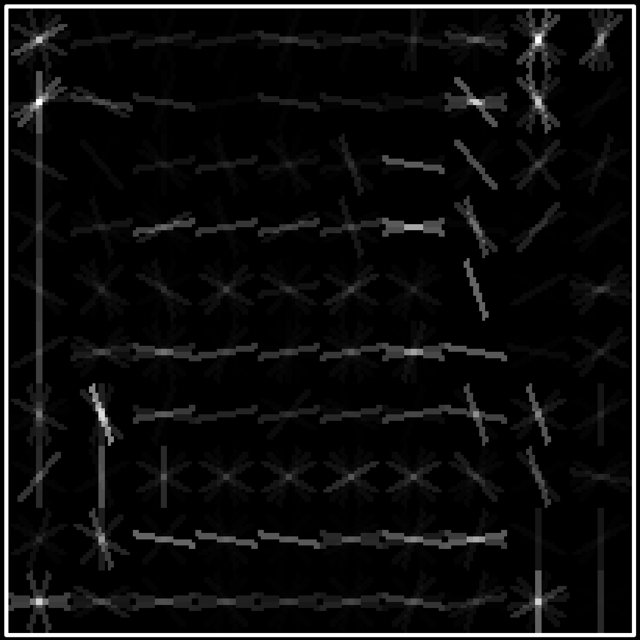


Style



Vanishing Points

Solving for 3D Structure

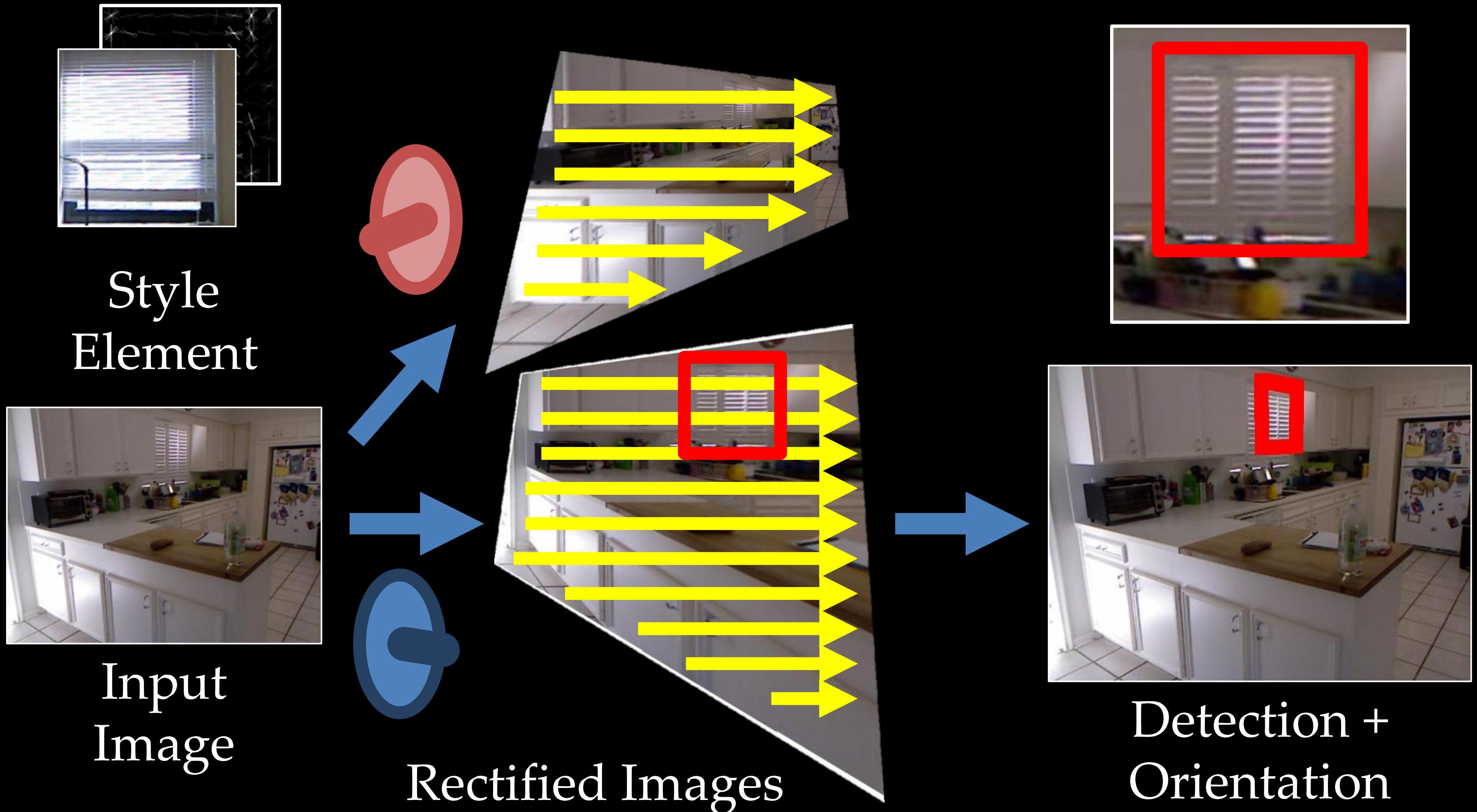


Style
Element

Input
Image

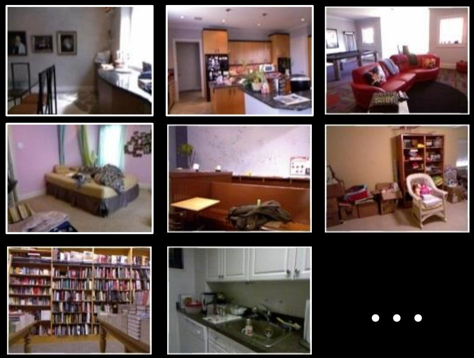
HOG, Dalal and Triggs '05; ELDA from Hariharan et al. '12

Solving for 3D Structure

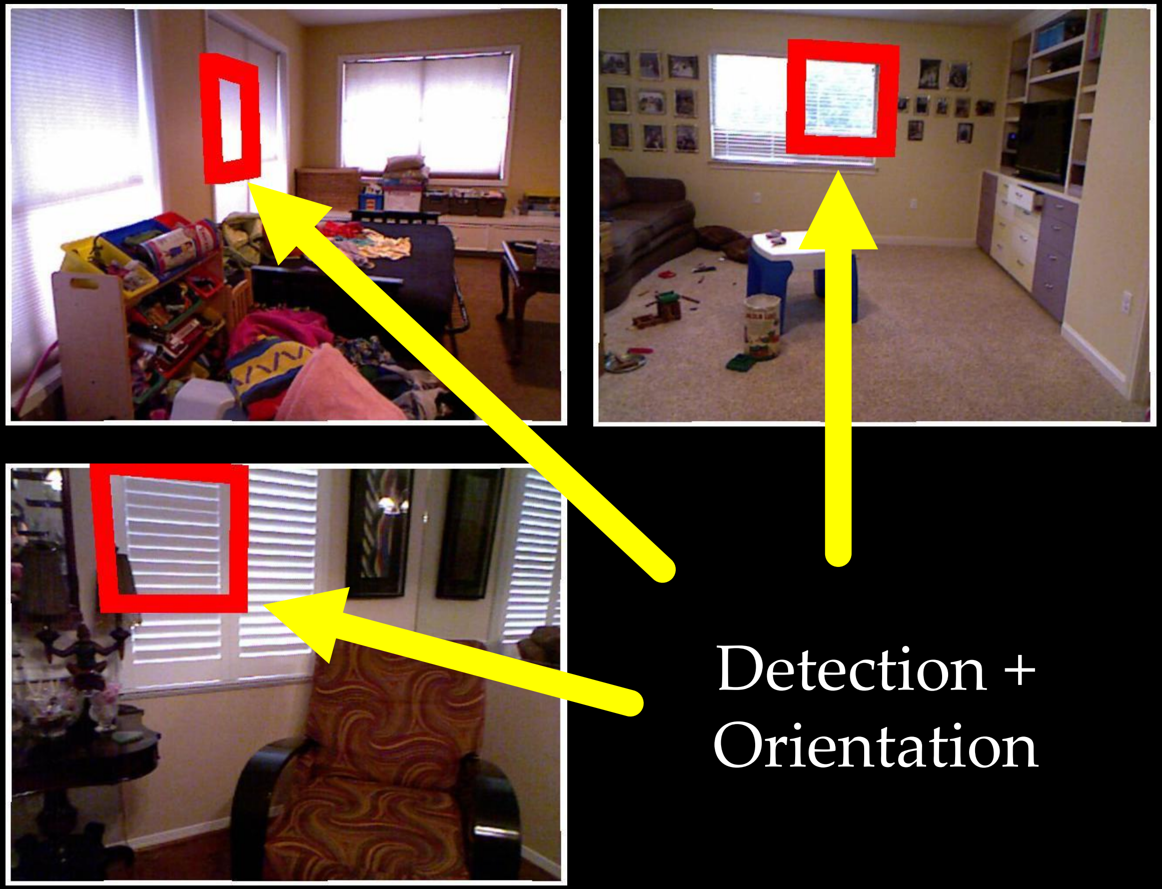


Solving for 3D Structure over a Dataset

Style
Element



Set of Images

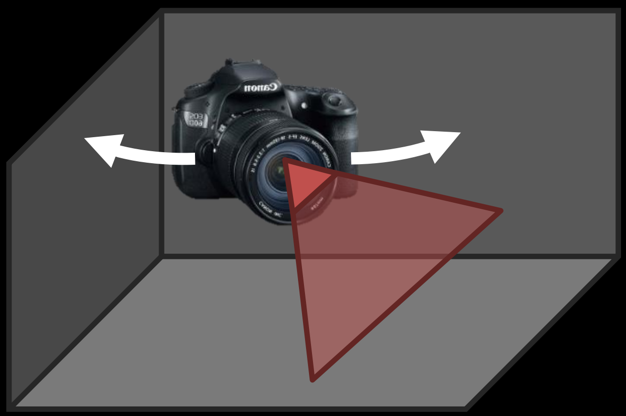


Detection +
Orientation

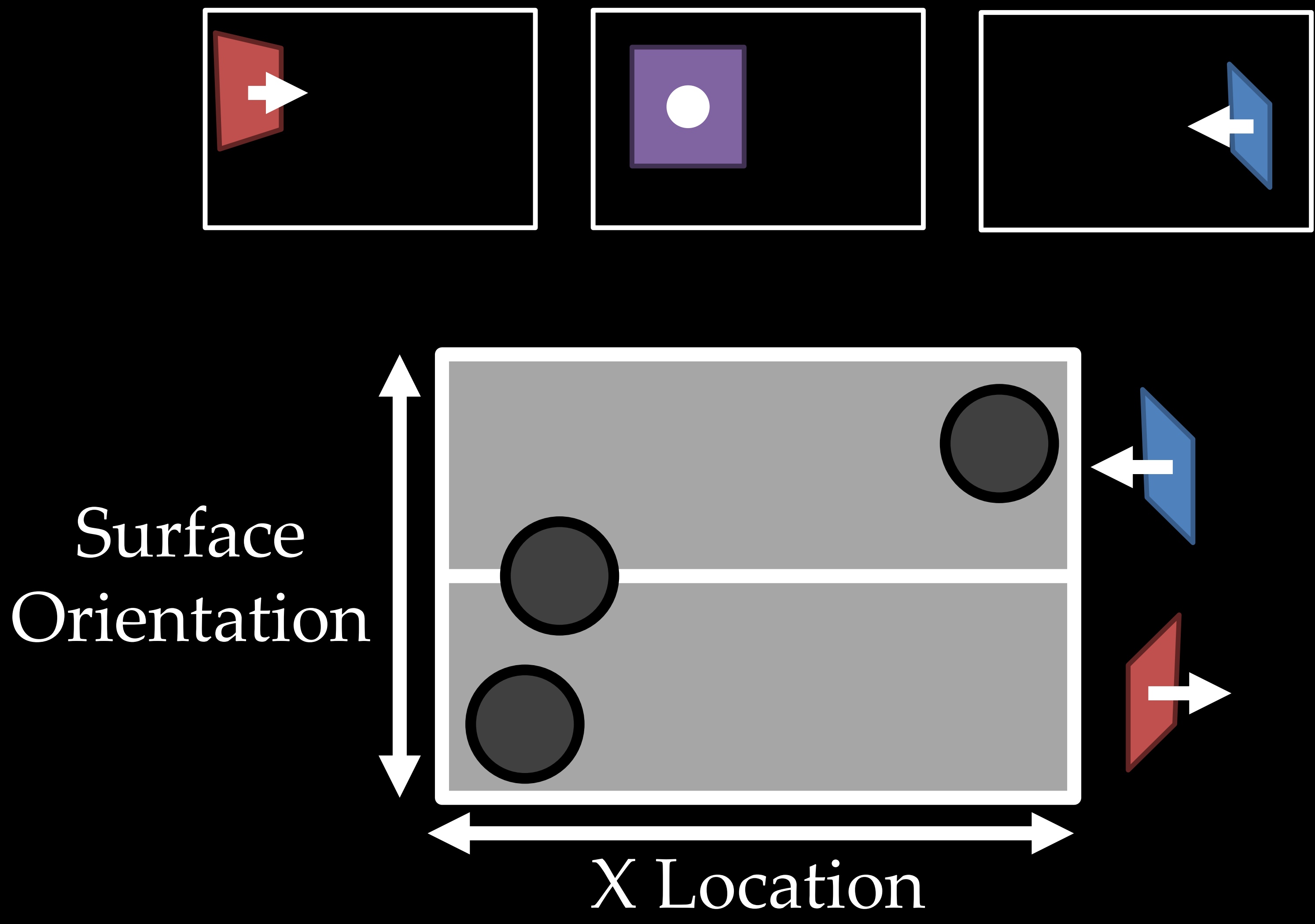
2 Key Assumptions

Style and 3D structure are independent

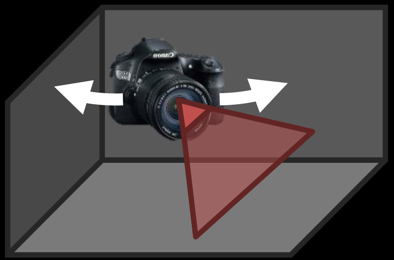
On average, 3D structure is a box

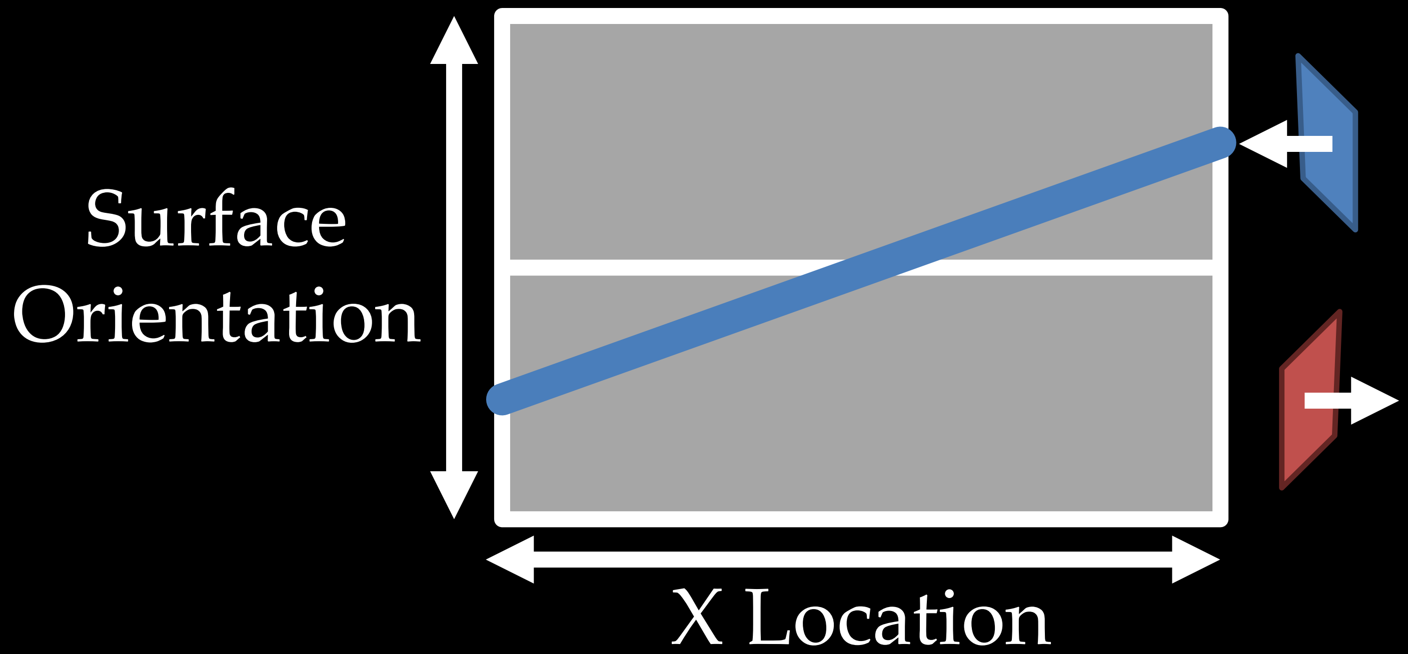


Plotting Detections

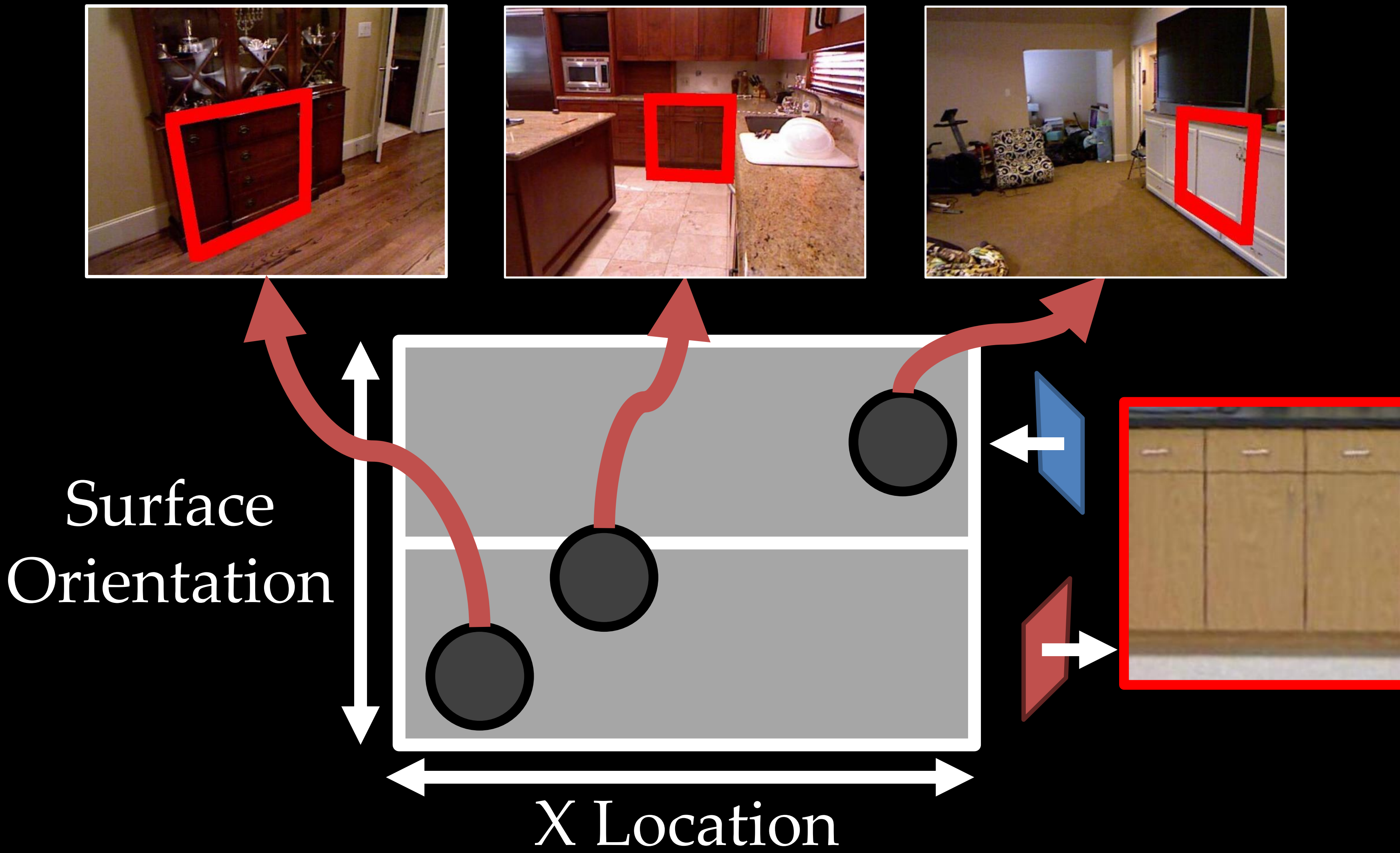


Box Assumption

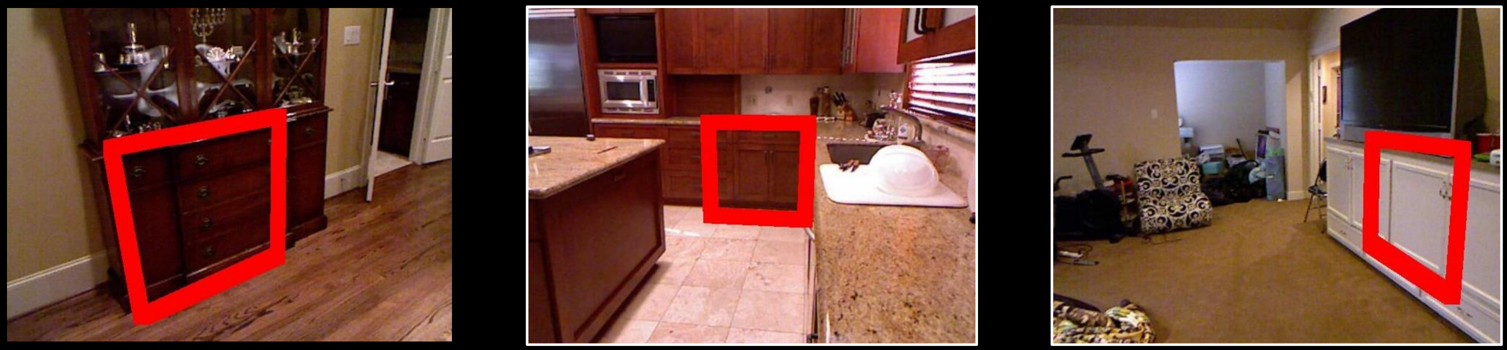


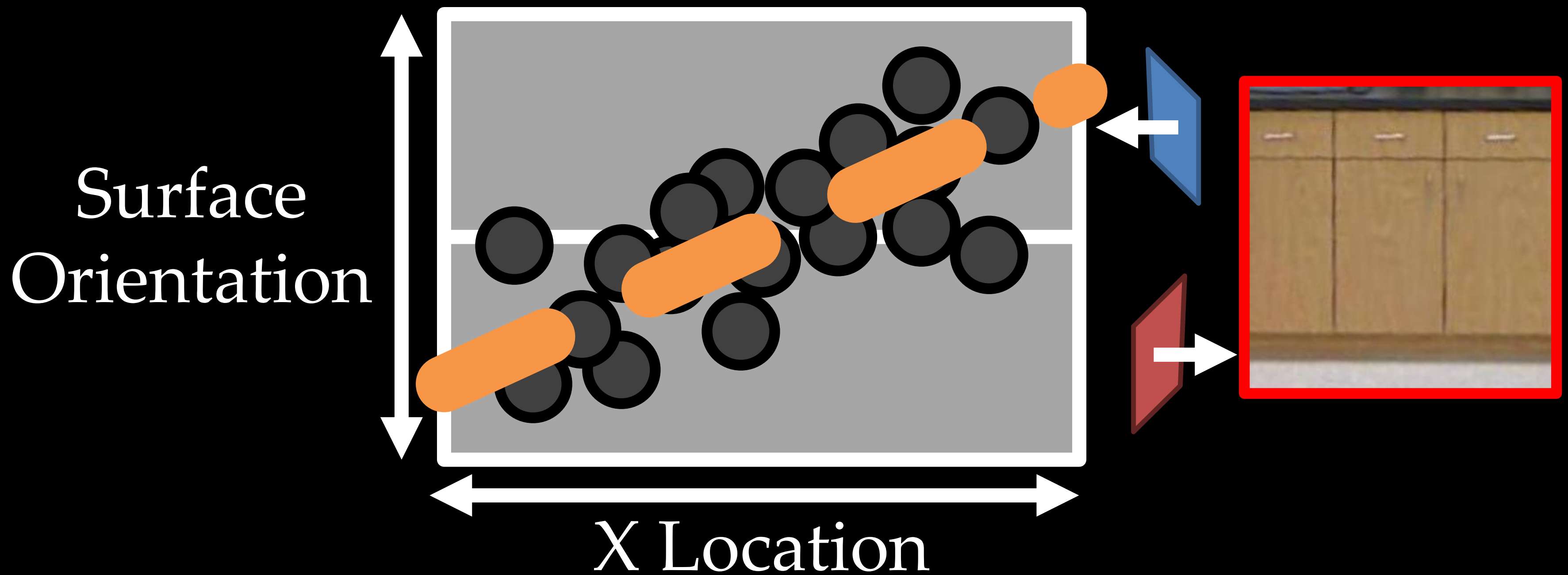


Verifying Style Elements

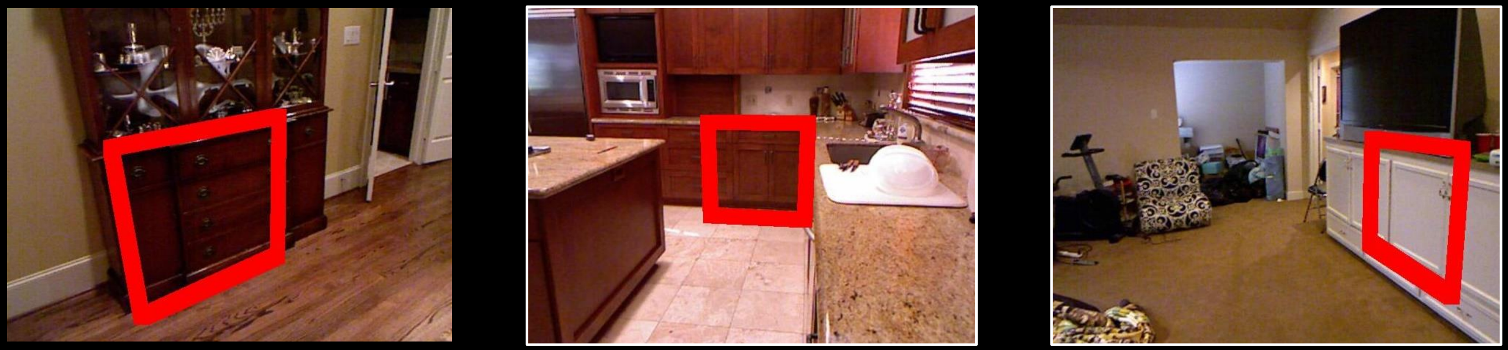


Verifying Style Elements

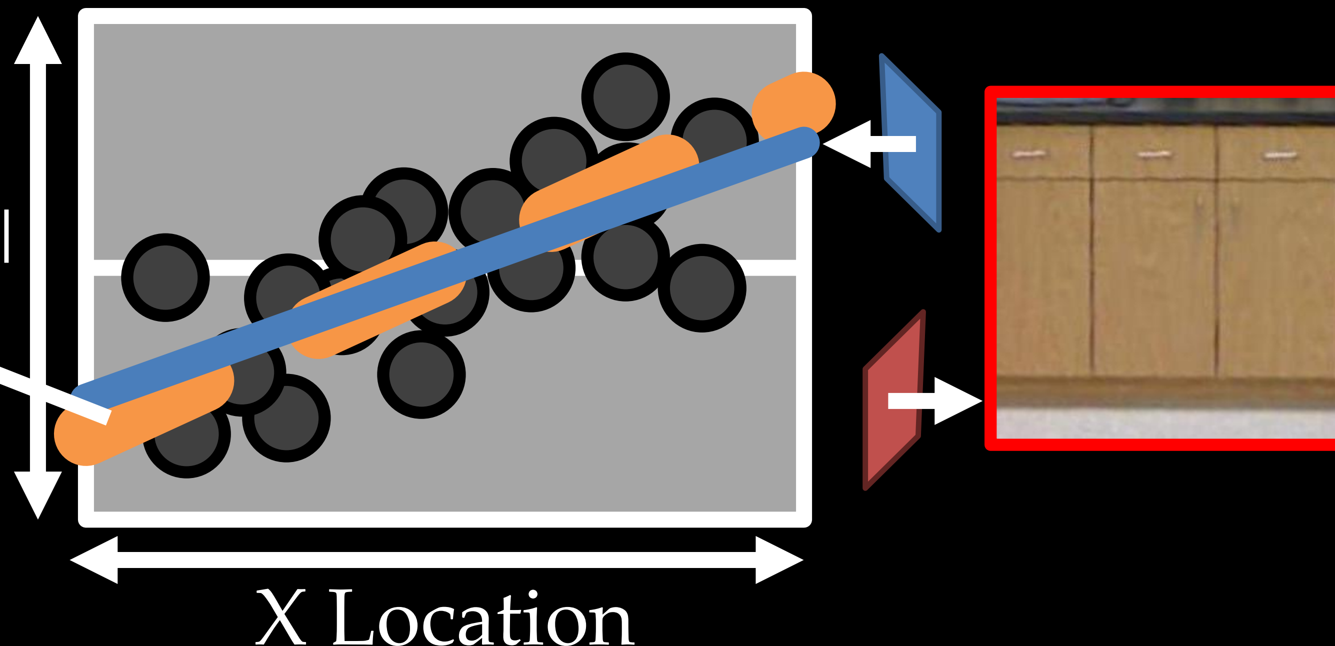




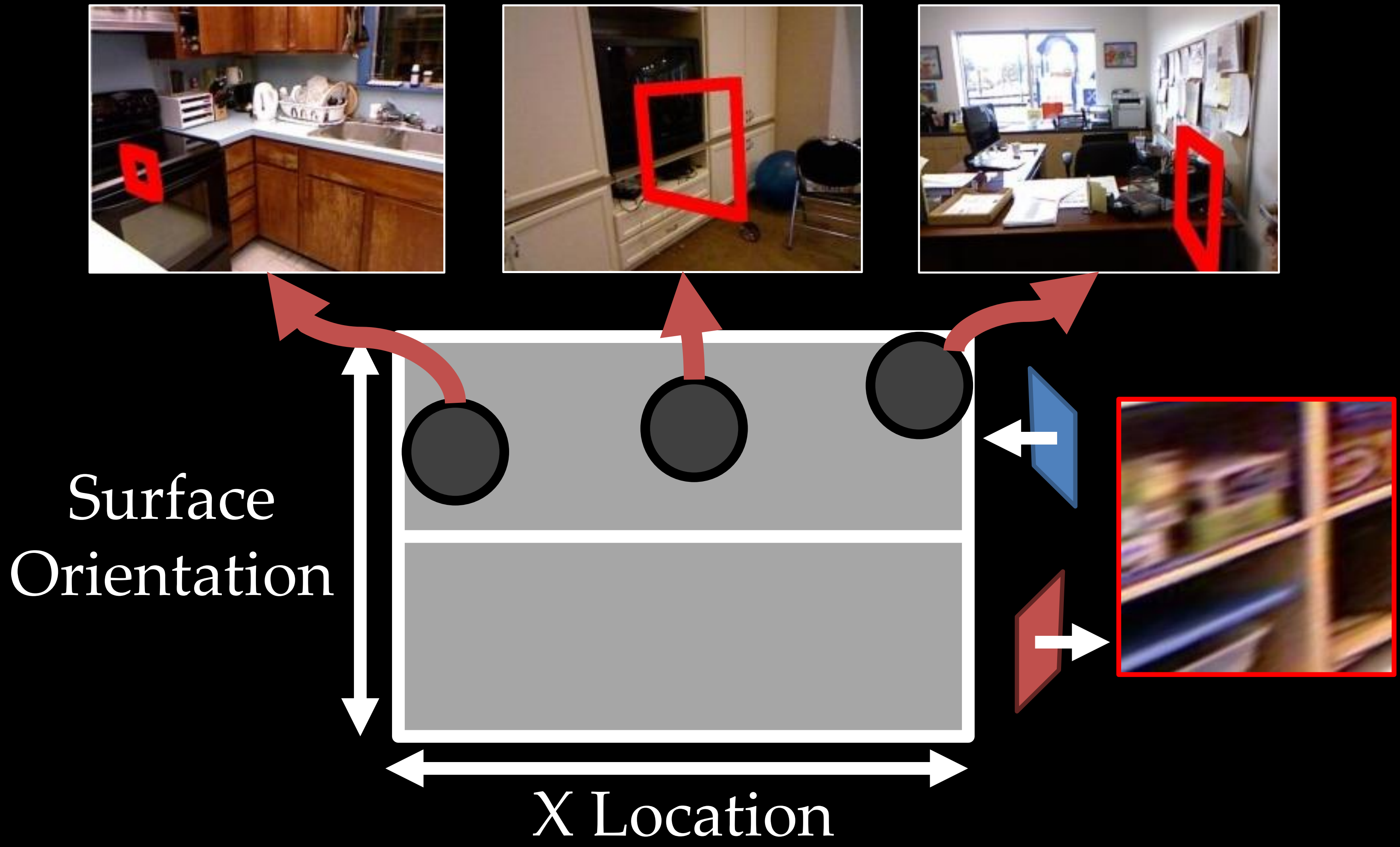
Verifying Style Elements



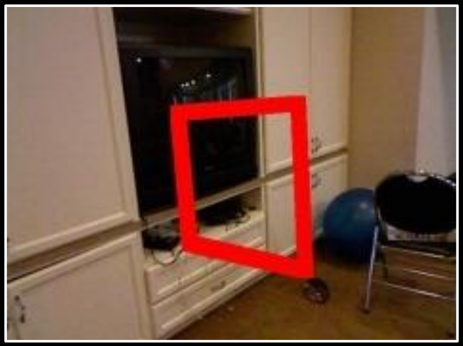
$$\sum_{i=1}^{W} |Prior_i - Data_i|$$

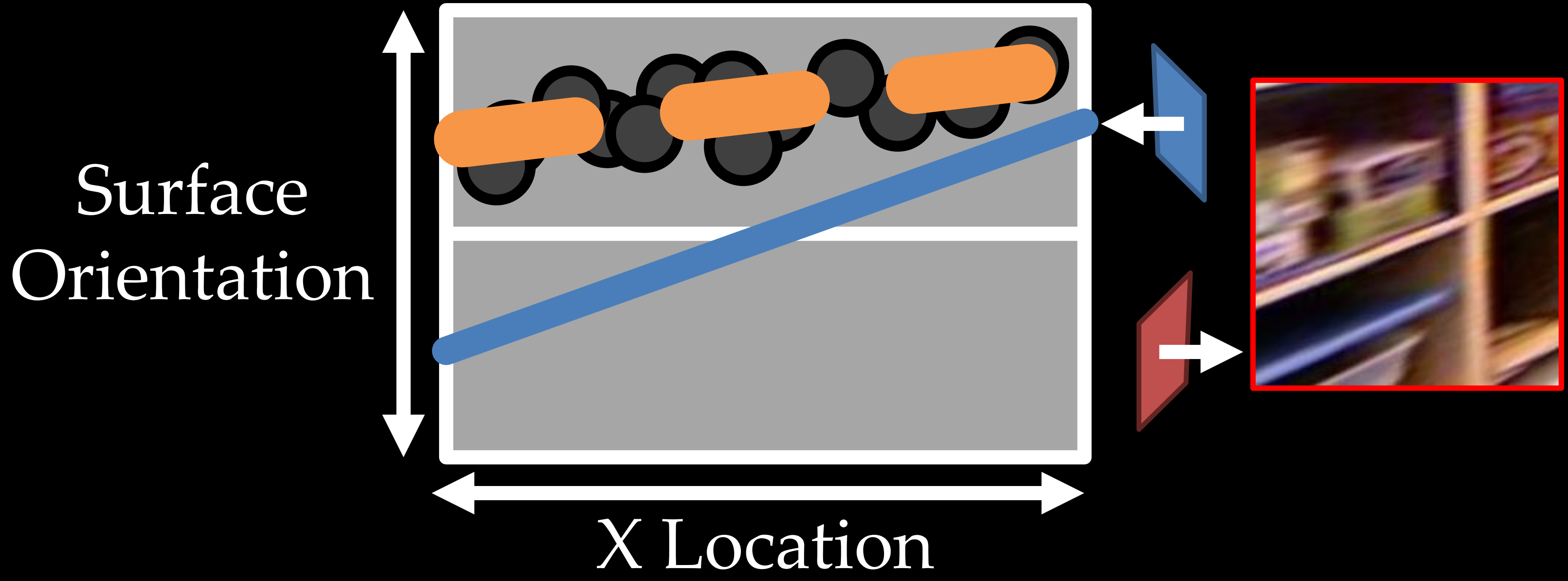


Verifying Style Elements

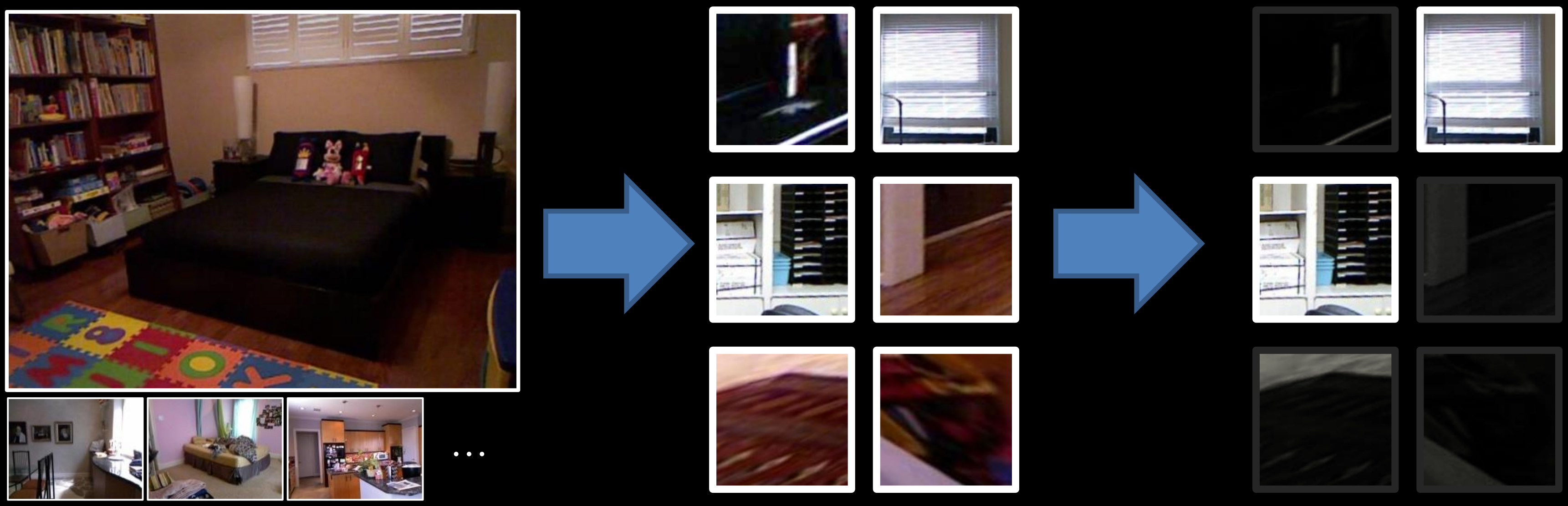


Verifying Style Elements



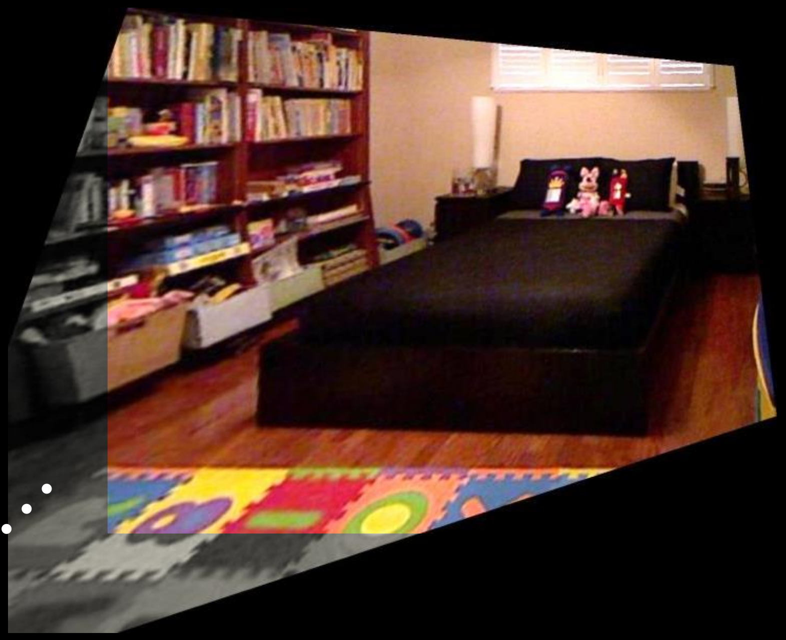


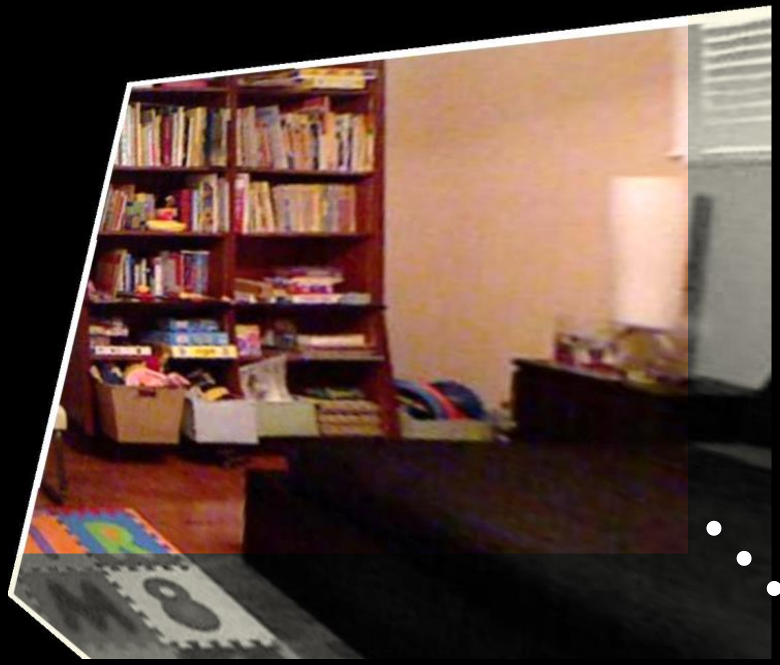
Hypothesize and Verify Pipeline



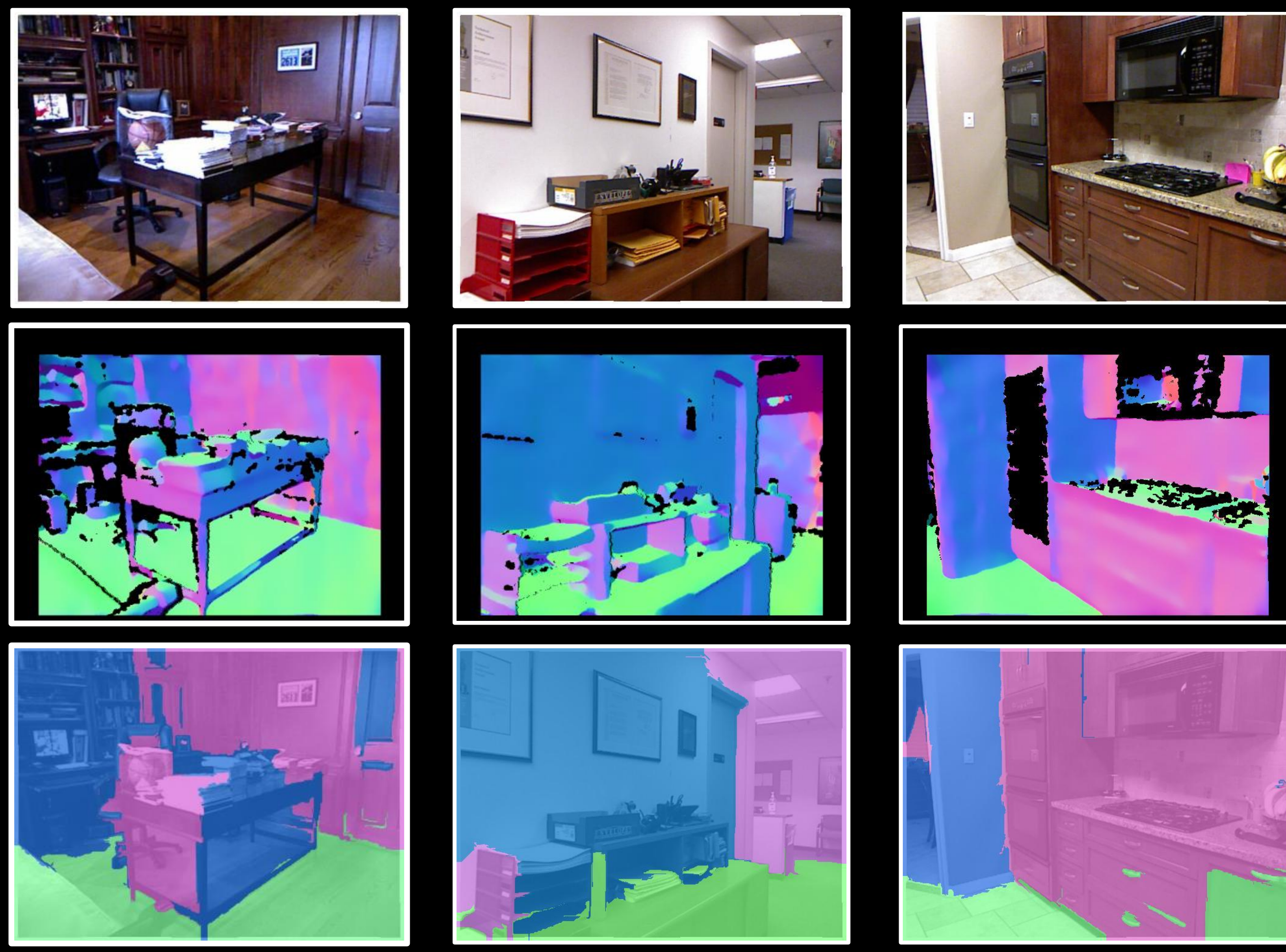
Discovered Style Elements

Interpreting





Results



Results

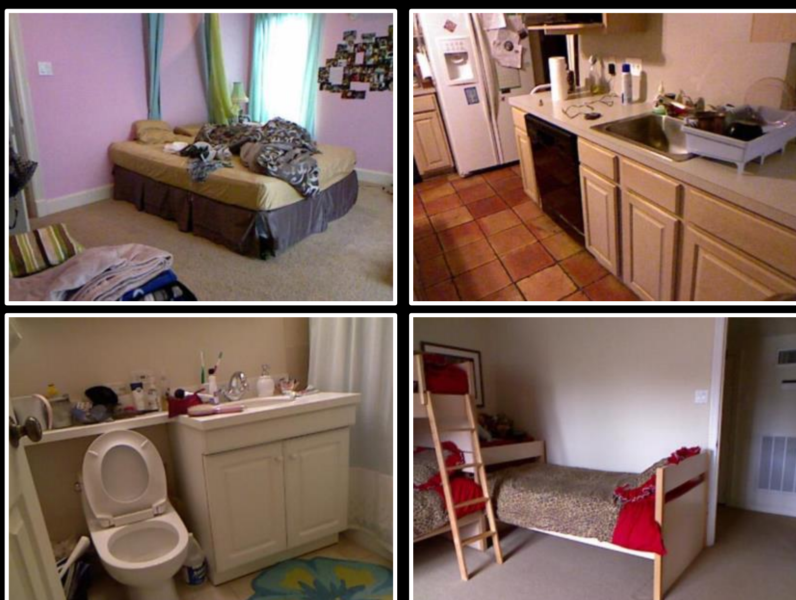
Quantitative Results

	All Pixels			Vertical
	(Lower Better) Median Error	(Higher Better) Pixels < 11.25°	Pixels < 30°	(Higher Better) Pixels < 30°
Style Elements	21.7°	36.8%	55.4%	59.7%
3DP	19.2°	39.2%	57.8%	58.8%
Origami World	17.9°	40.5%	58.9%	
Disc. Coding	23.5°	27.7%	58.7%	

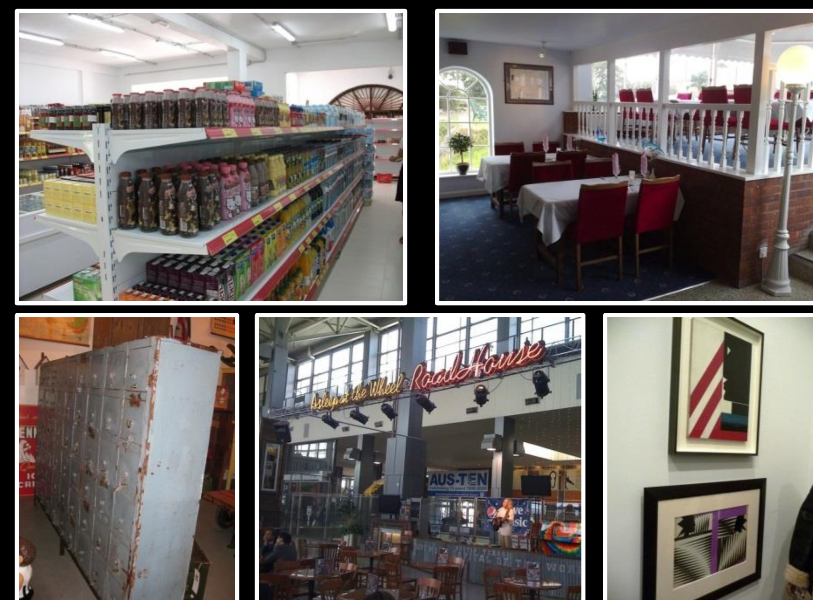
3DP: Fouhey et al. ICCV '13; Origami World: Fouhey et al. ECCV '14; Disc. Coding: Ladicky et al. ECCV '14

Scaling Up To The World

RGBD Datasets



Internet Images



Images from Places-205, Zhou et al. NPS '15

Results on Internet Images

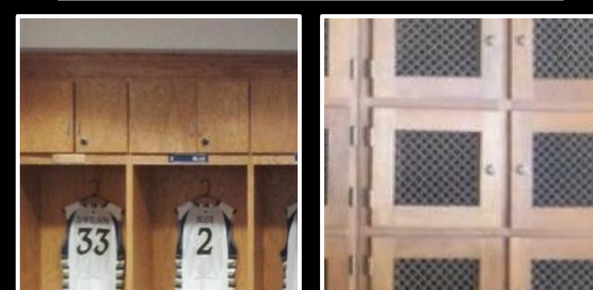
Automatically Discovered Style Elements

Supermarket

Laundromat

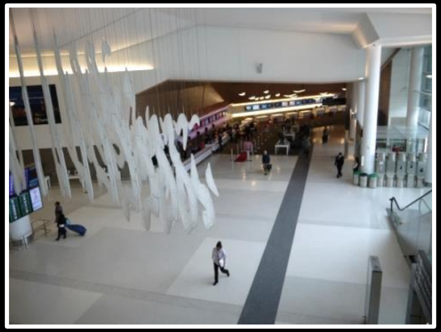
Museum

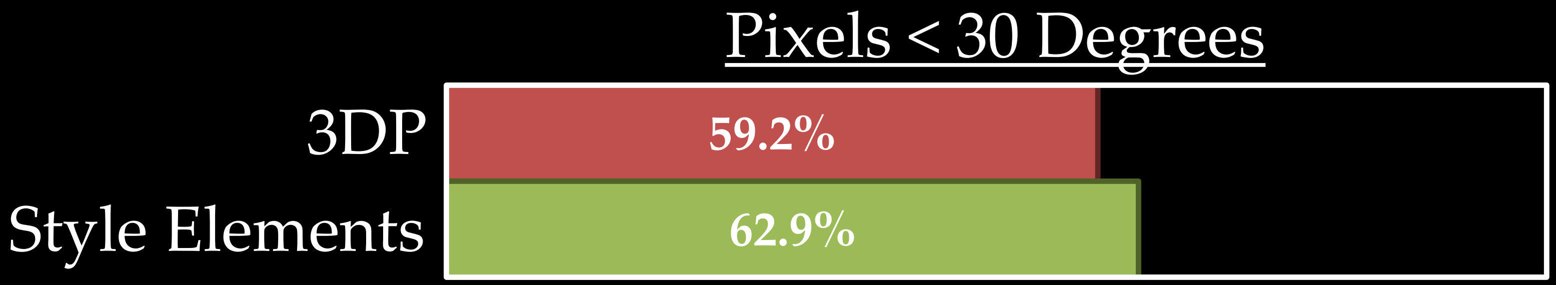
Locker Room



Quantitative Results

10 categories from Places-205 Dataset
Images sparsely manually annotated



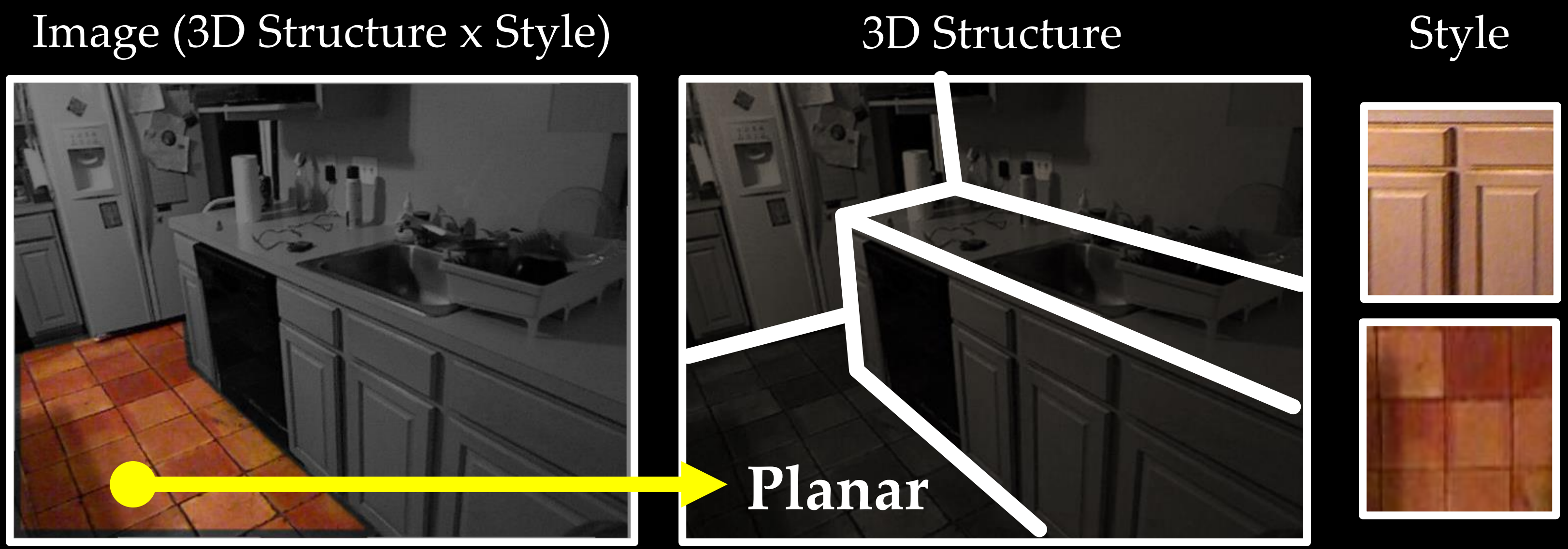


3DP: Fouhey et al. ICCV '13. Images from Places-205, Zhou et al. NPS '15

The Story So Far

Unconstrained Outputs

Constrained Outputs



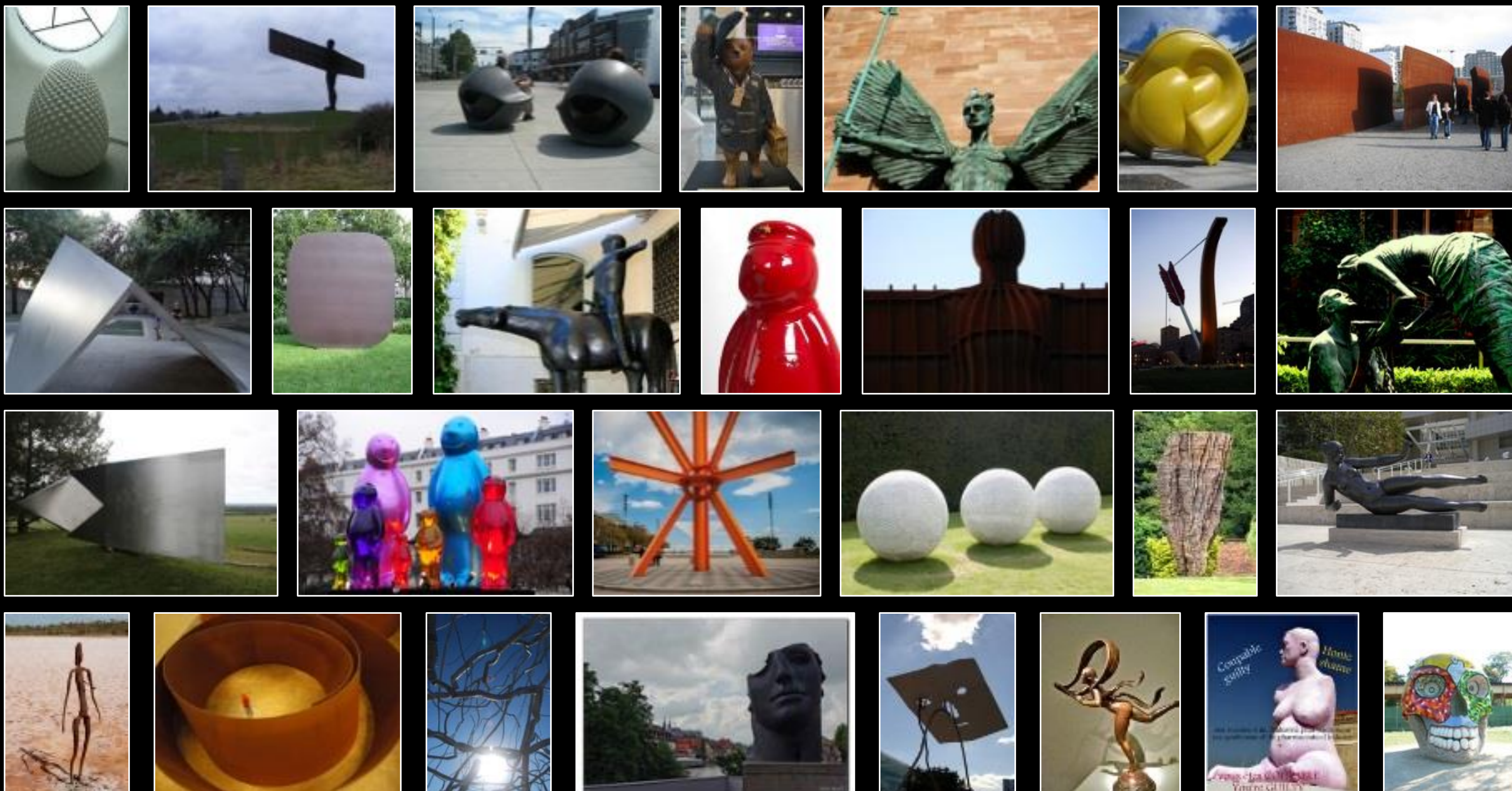
Cues for higher-order 3D structure

3D SHAPE ATTRIBUTES

Goal: 3D Shape Attributes

Not Planar
Smooth surface
1 point of contact
Not point contact
Has Hole
Not thin structures
…

Data



flickr

3D Shape Attributes

Curvature
(4 Total)

Planar Surfaces

Cylindrical Surfaces

Contact
(2 Total)

Point or Line

Multiple

Occupancy
(6 Total)

Thin Structures

Has Hole

Examples

Positives: Has Planar Surfaces



Examples

Negatives: Has Planar Surfaces

Examples

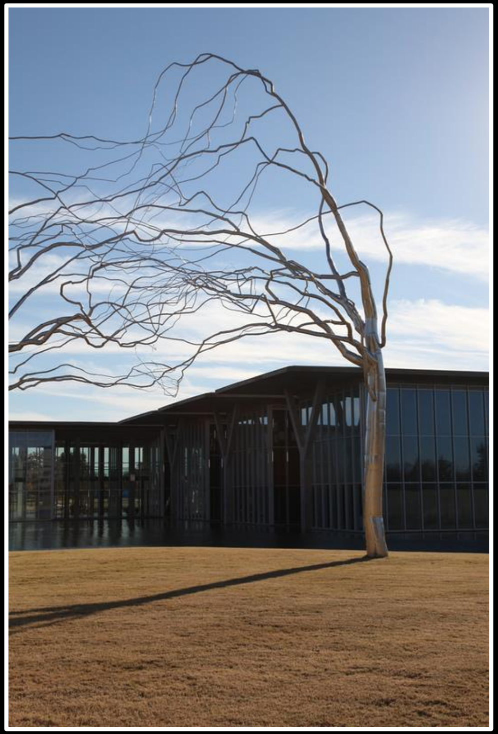
Positives: Has Point/Line Contact

Examples

Negatives: Has Point/Line Contact

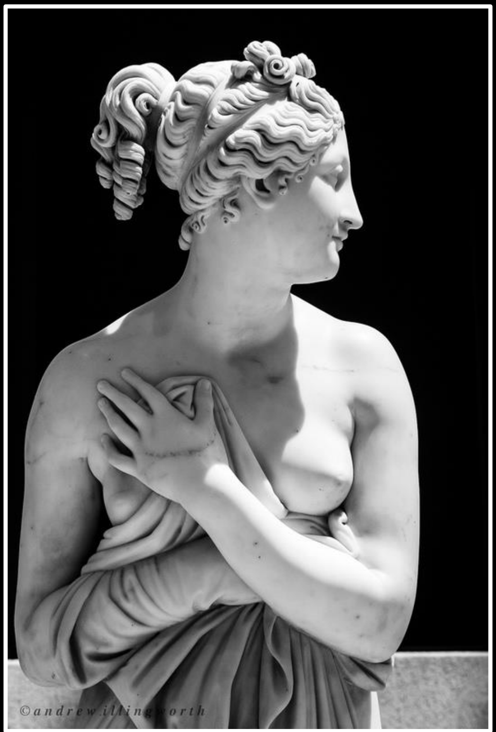
Examples

Positives: Has Thin Structures


Examples

Negatives: Has Thin Structures



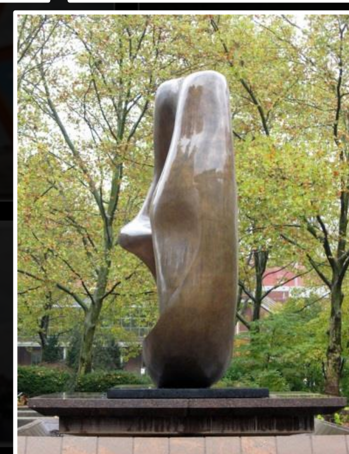
Data

London

Malaga

Yorkshire

Princeton

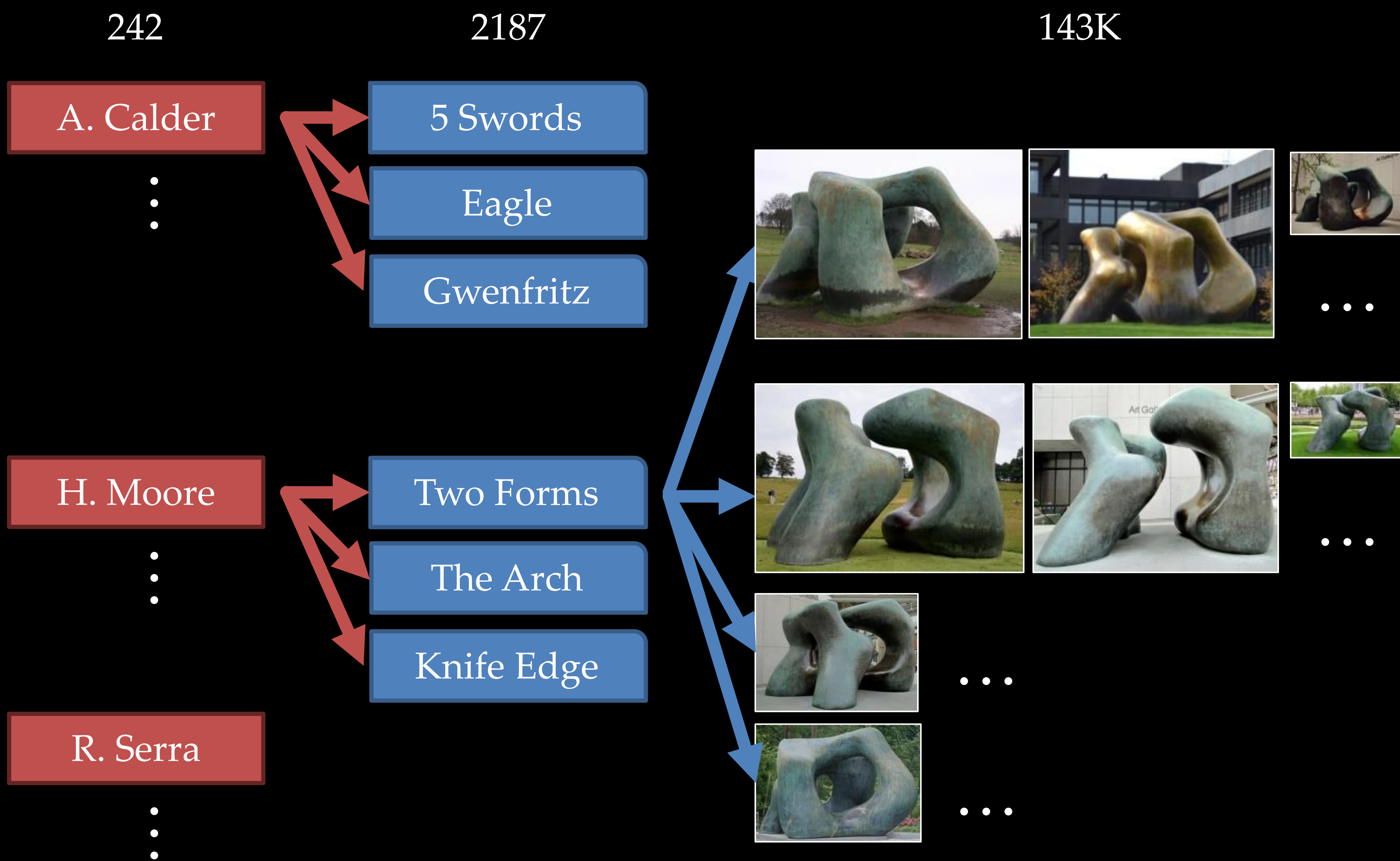


Columbus

Toronto

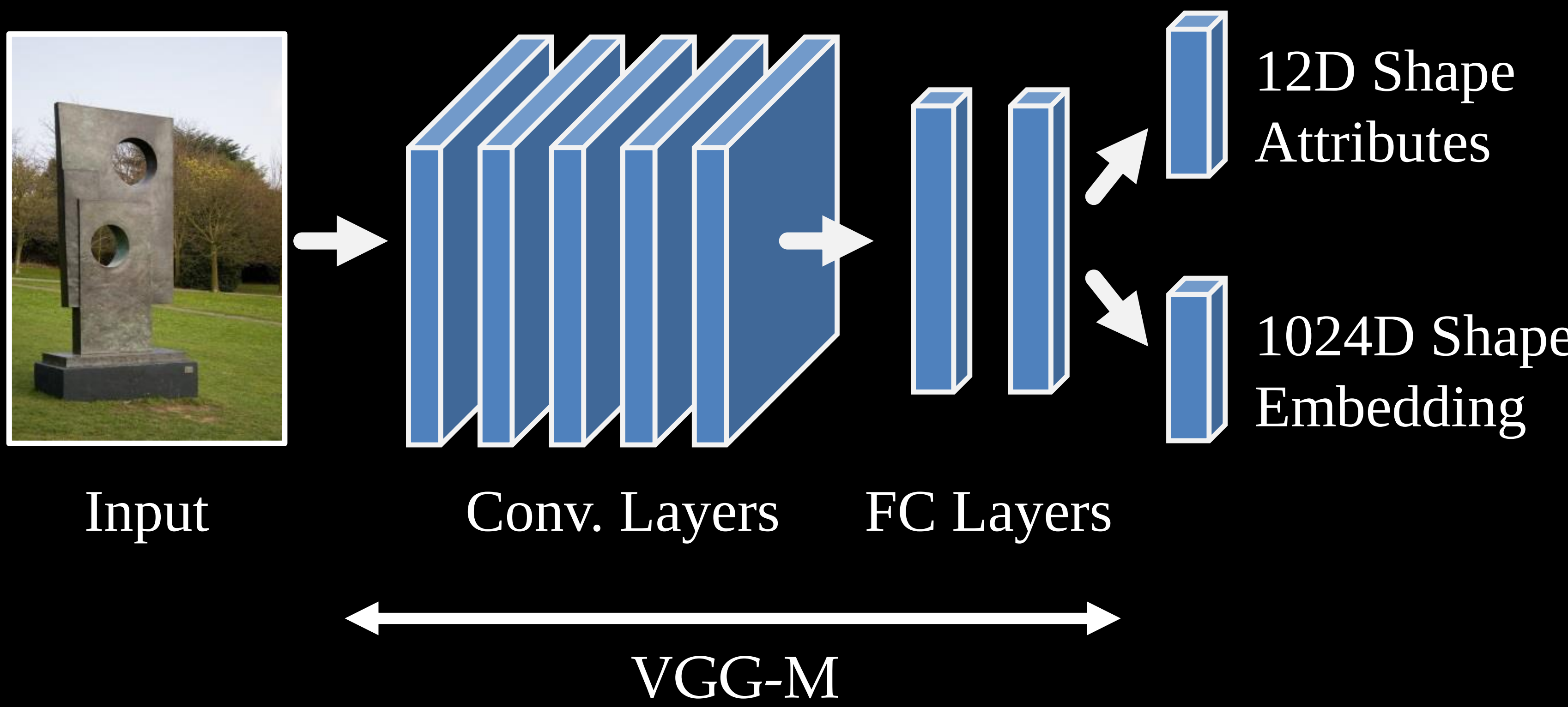
Data

Data



242
2187
143K
A. Calder
5 Swords
Eagle
Gwenfritz
H. Moore
Two Forms
The Arch
Knife Edge
R. Serra

Learning To Predict



Triplet loss as in Schults and Joachims ’04, Schroff et al. ’14, Wang et al. ‘15, Parkhi et al. ‘15

Qualitative Results

Most ← Point/Line Contact → Least


...

Rough Surface


...

Indirect Baselines

Planar = Yes
Holes = Yes
…
2+ Contacts = No
…

- SIRFS (Barron et al. ’15)
- CNN (Eigen et al. ’14)
- KDES+SVM (Bo et al. ‘11)
- HHA+CNN (Gupta et al. ‘14)

Quantitative Results

Criterion: mean AUC of ROC.

Eigen ‘14		Barron ‘15		End-to-end
KDES	HHA	KDES	HHA	
58.5	61.2	59.4	62.5	**72.3**

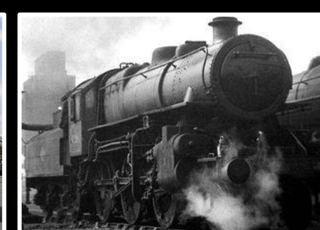
PASCAL VOC Results

Most Planarity Least

. . .

Most Planarity Least

. . .



PASCAL VOC Results

Most ← Rough Surface → Least


...


Most ← Point/Line Contact → Least


...



The Story So Far

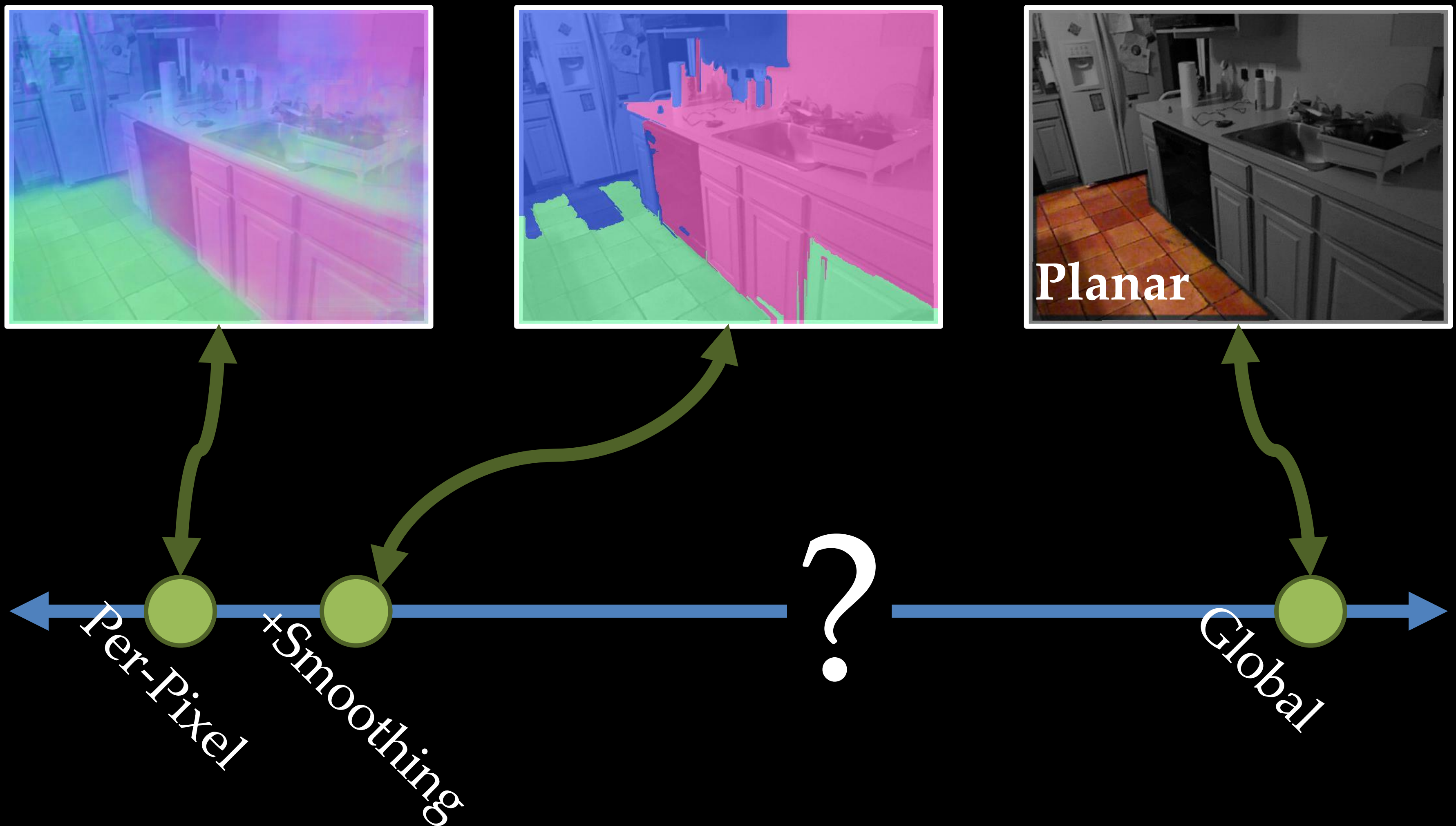
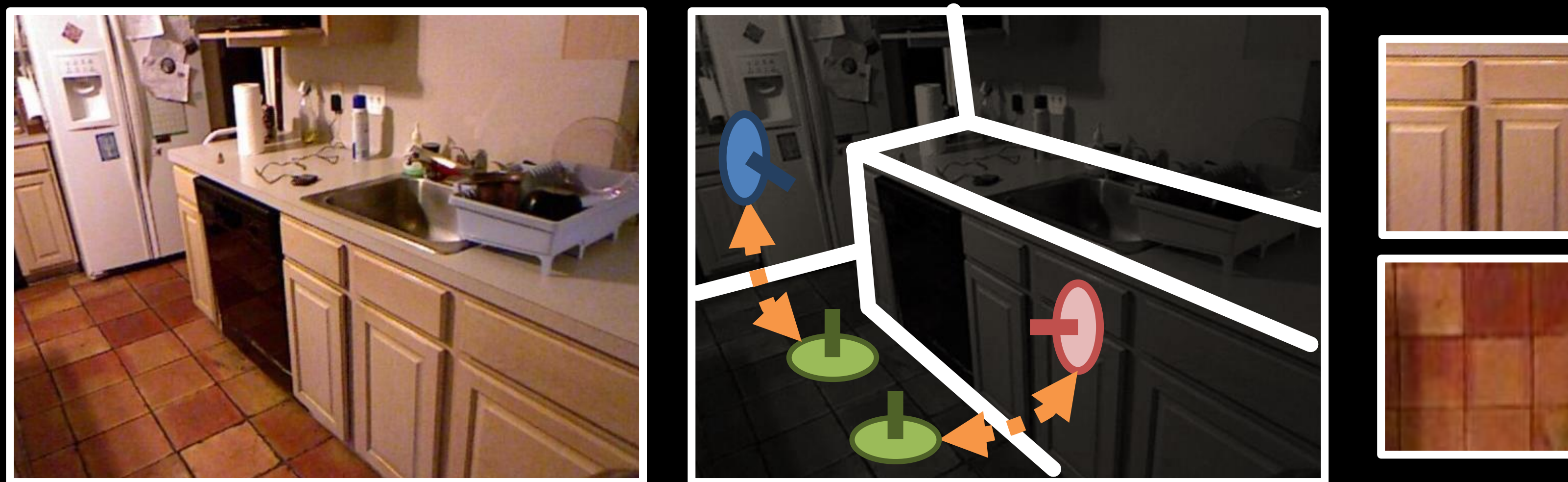


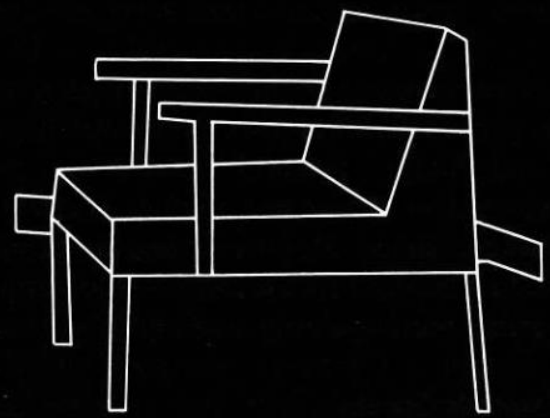
Image (3D Structure x Style) 3D Structure Style

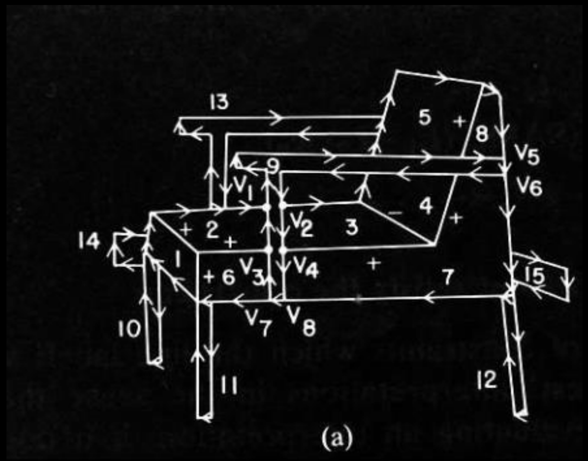


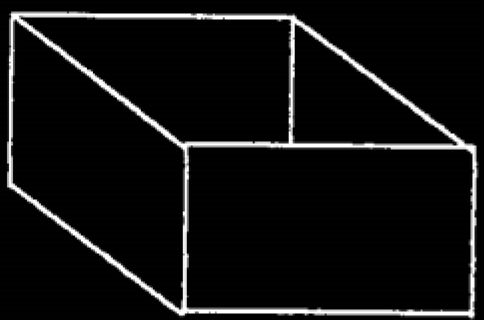
Mid-level constraints on 3D Structure

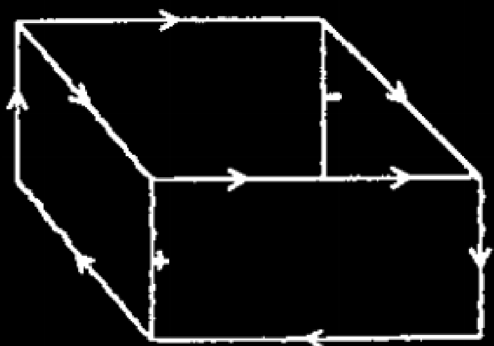
CONSTRAINTS ON 3D STRUCTURE

Mid-level in the Past



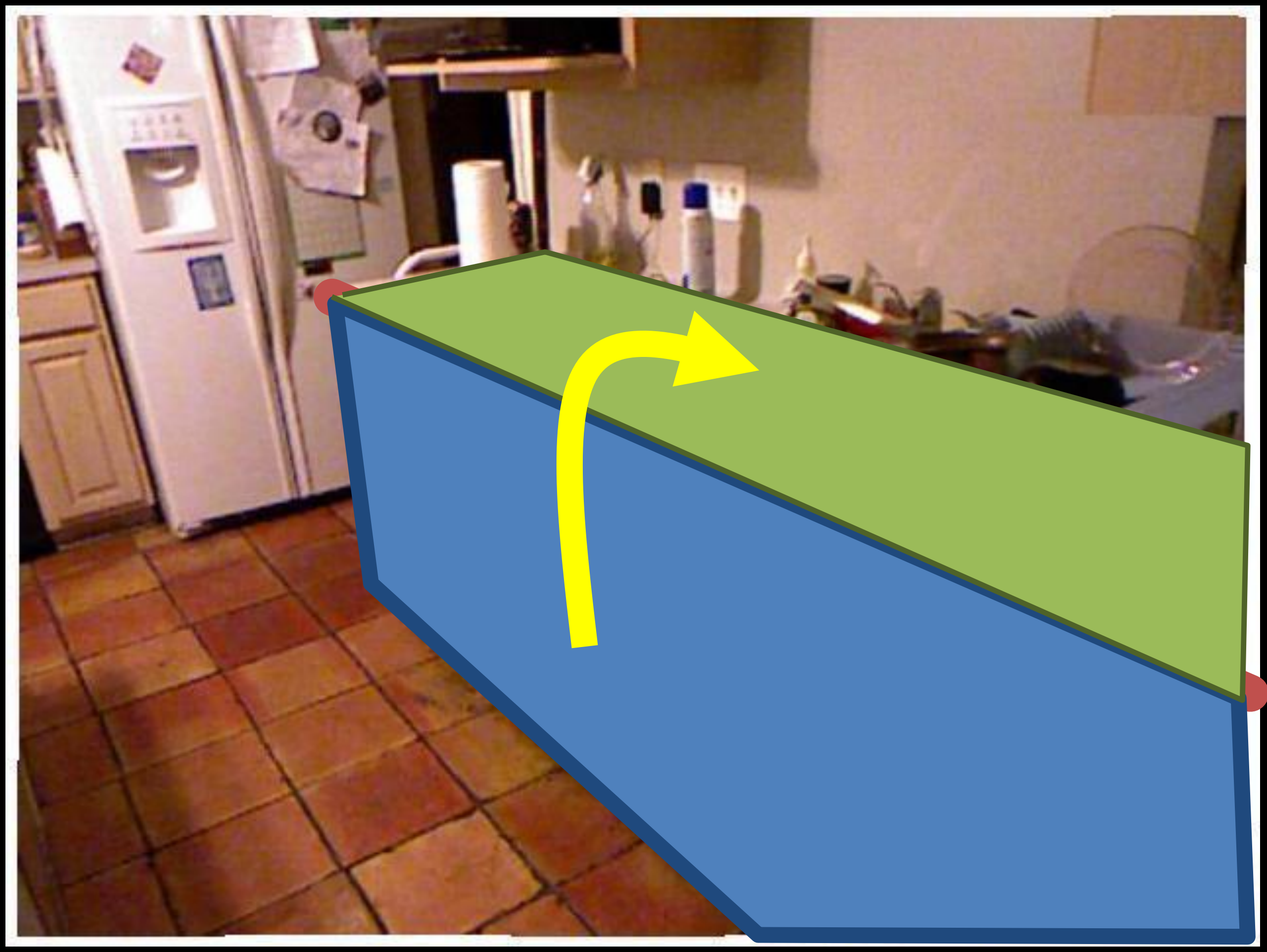






Huffman 71, Clowes 71, Kanade 80, 81 Sugihara 86, Malik 87, etc.

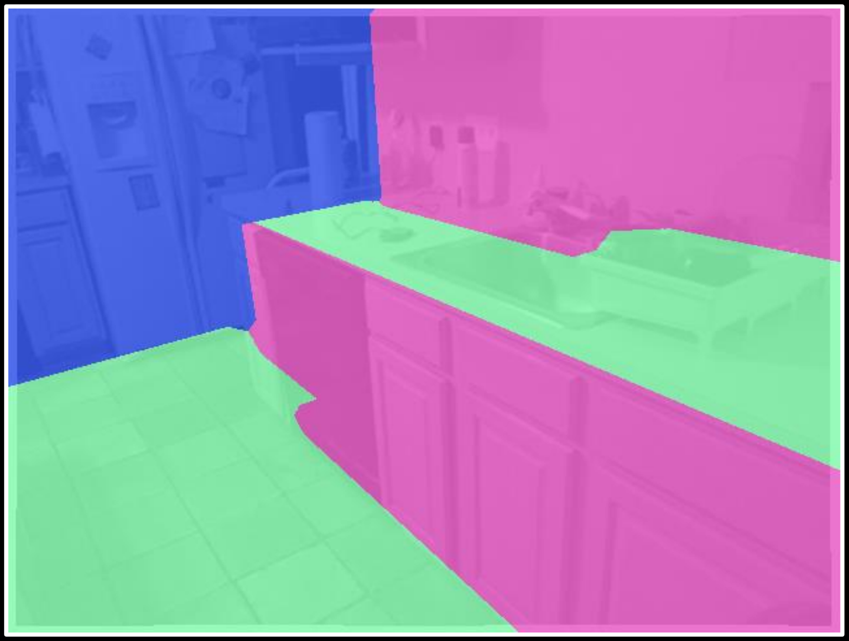
Our Mid-Level Constraints



Our Output

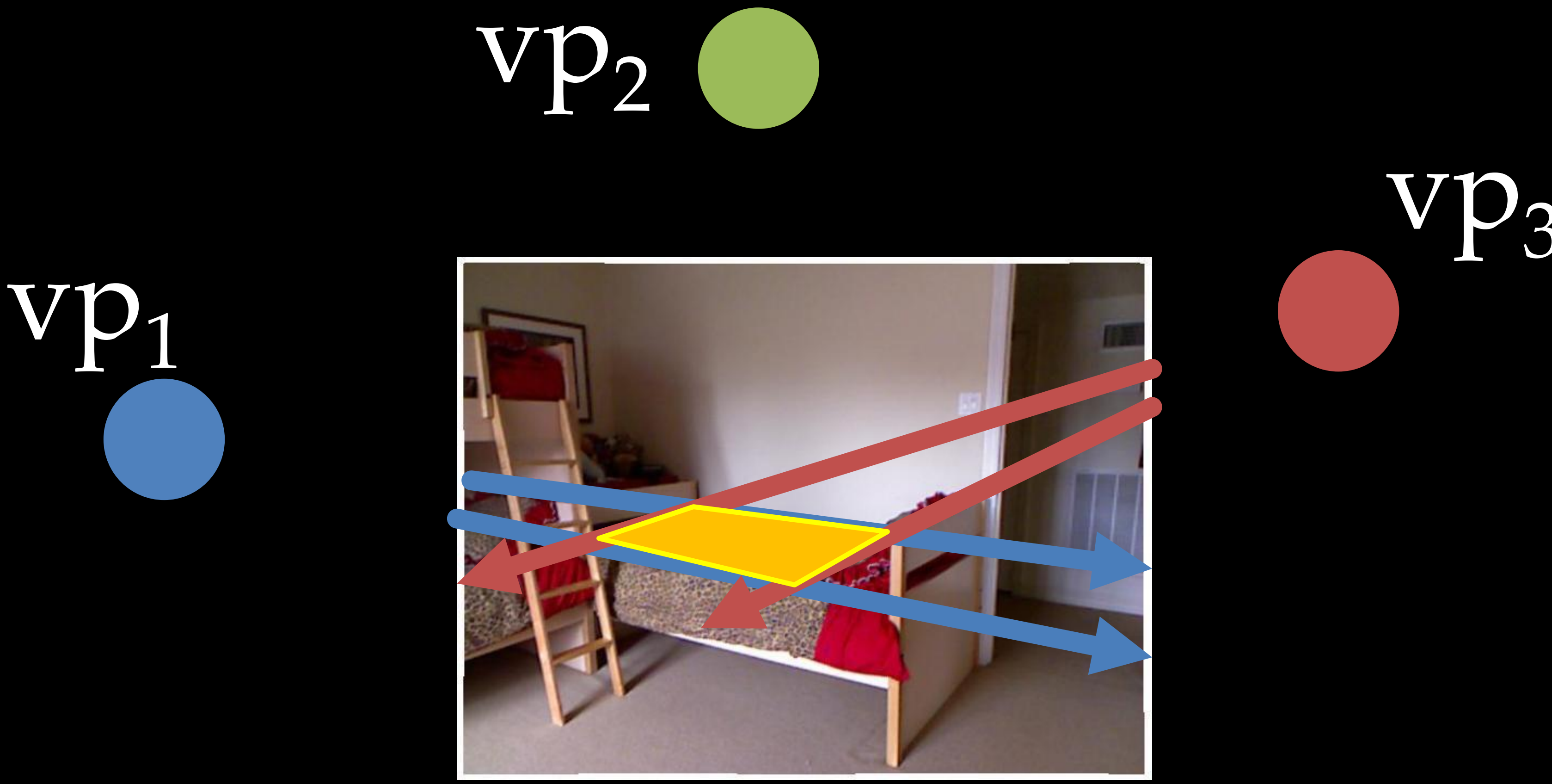
Input:
Single Image

Output:
Discrete Scene Parse



Parameterization

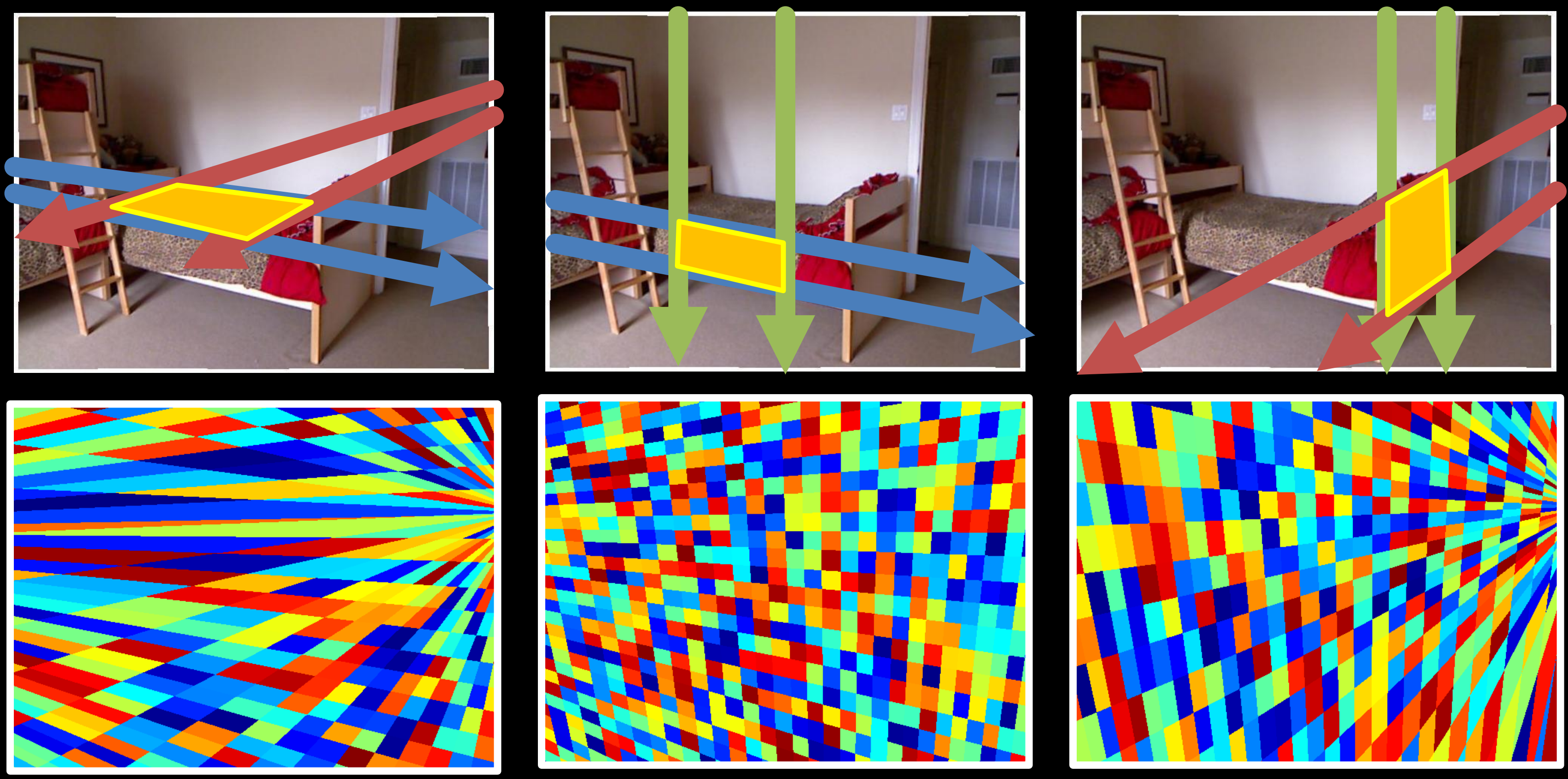
Parameterization



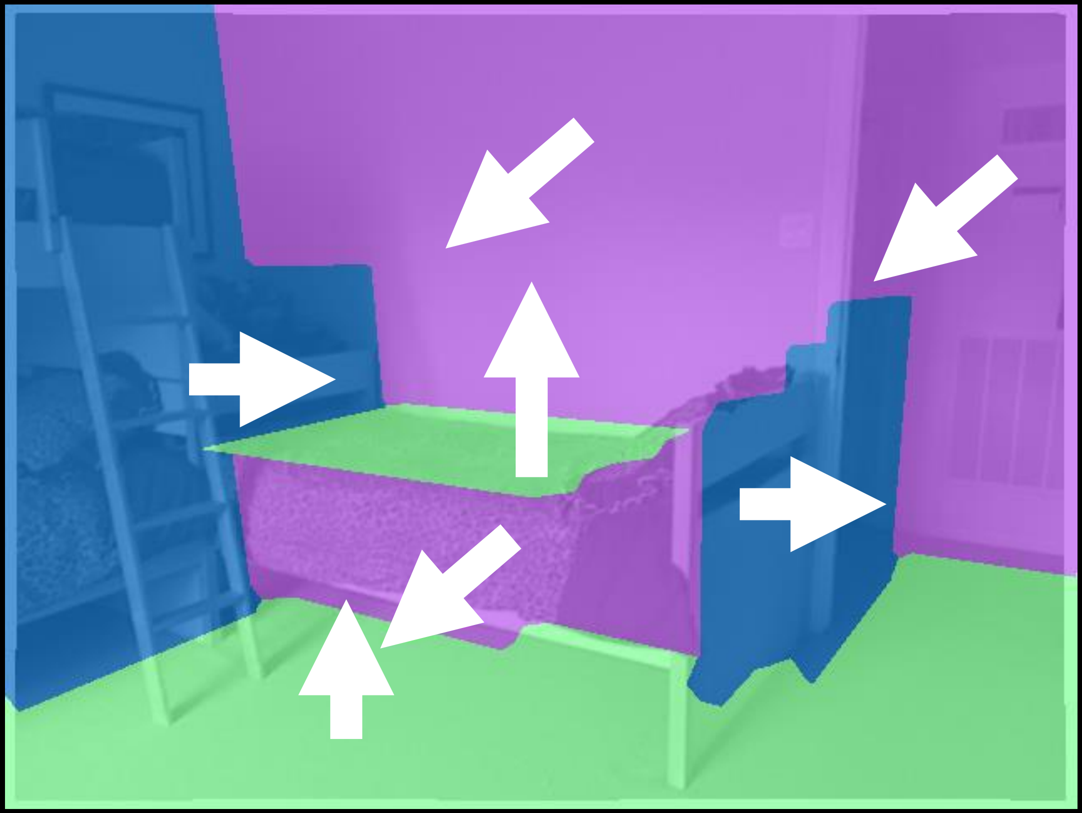
VP Estimator from Hedau et al., 2009

Parameterization

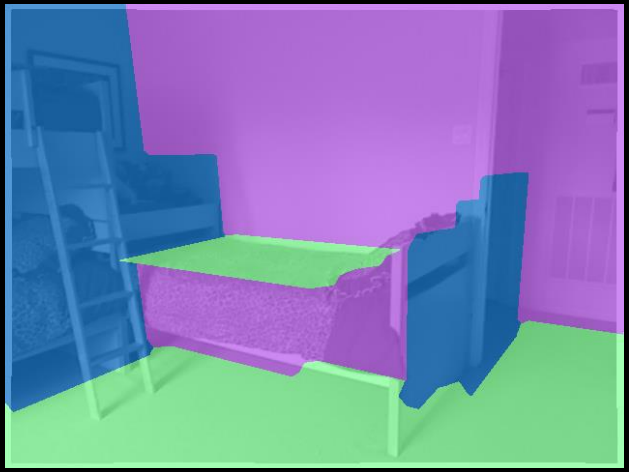
Two VPs give grid cell

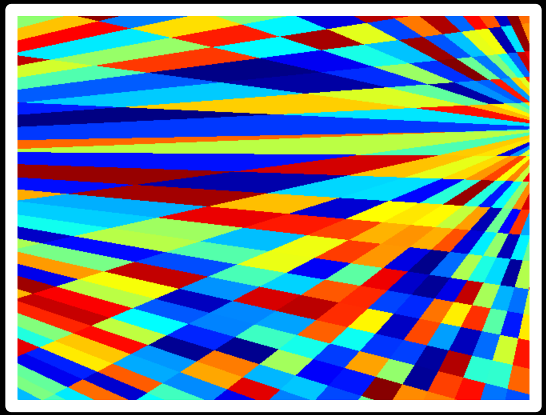


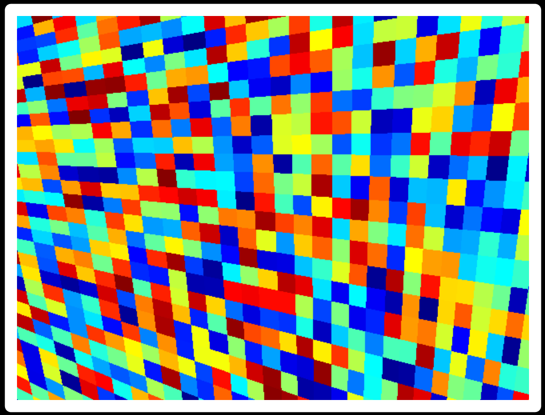
Encoding Surface Normals

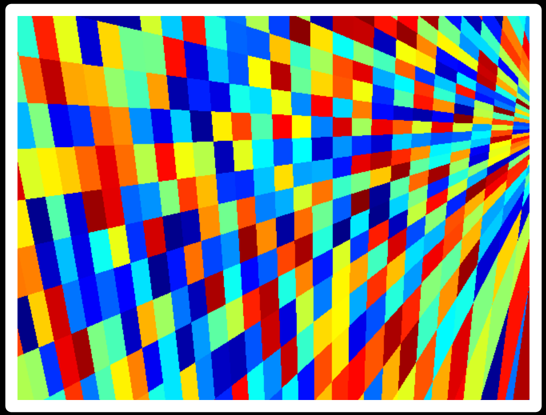


Encoding Surface Normals

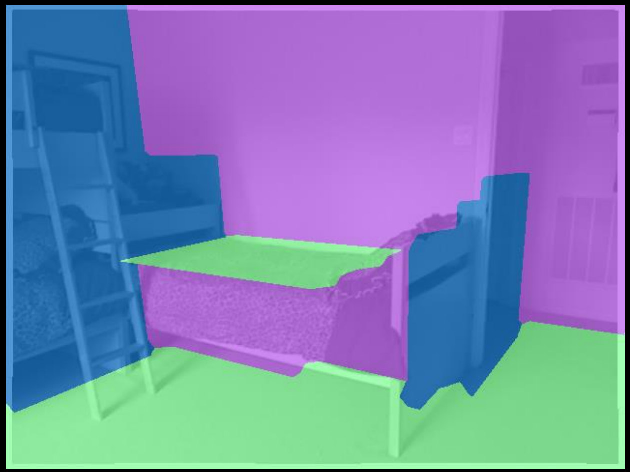


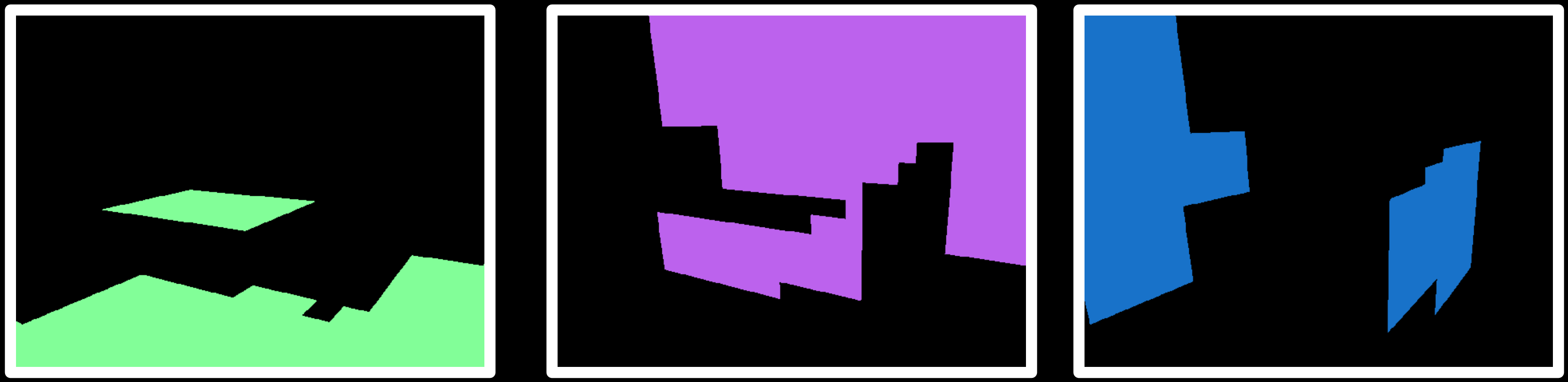




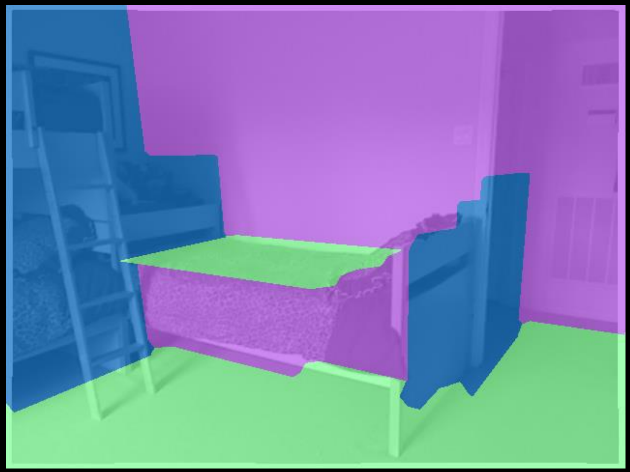


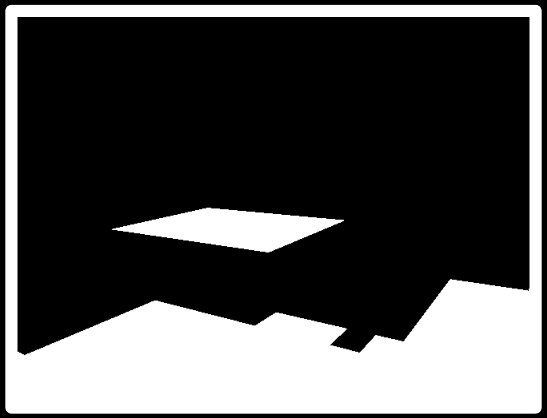
Encoding Surface Normals



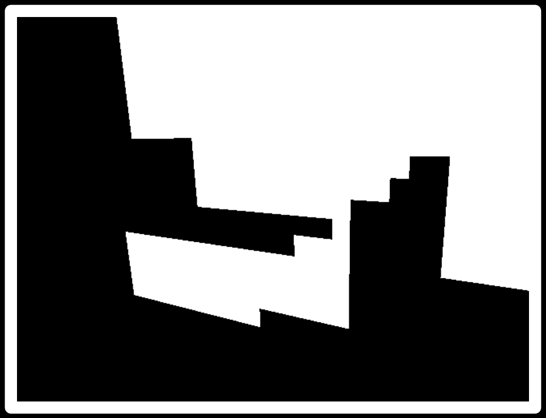


Encoding Surface Normals

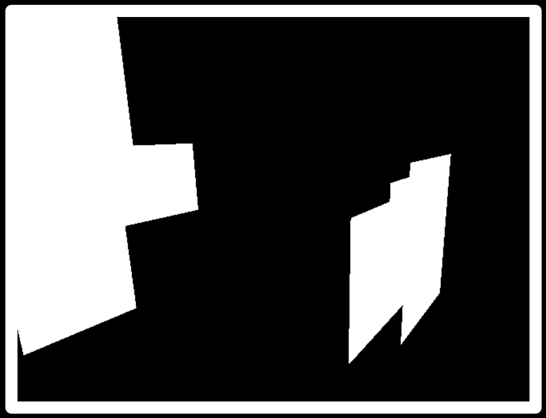




$x_1, \ldots, x_{400}$



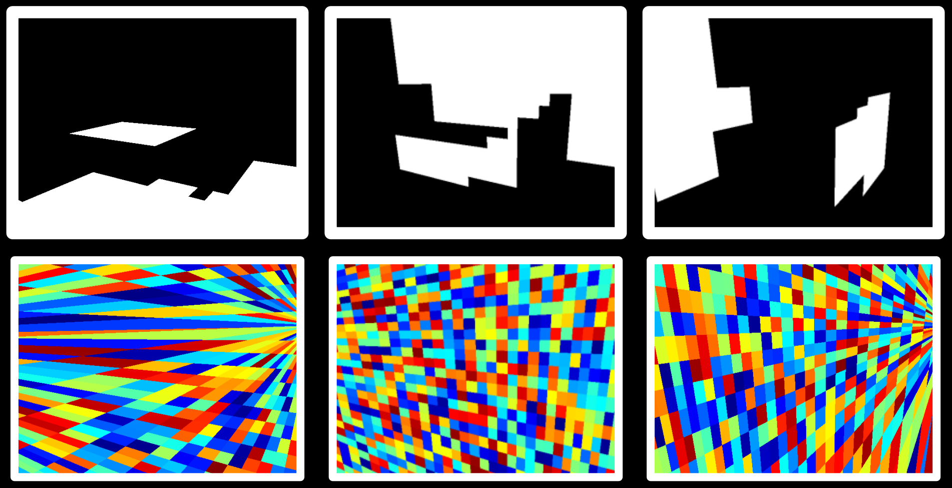
$x_{401}, \ldots, x_{800}$



$x_{801}, \ldots, x_{1200}$

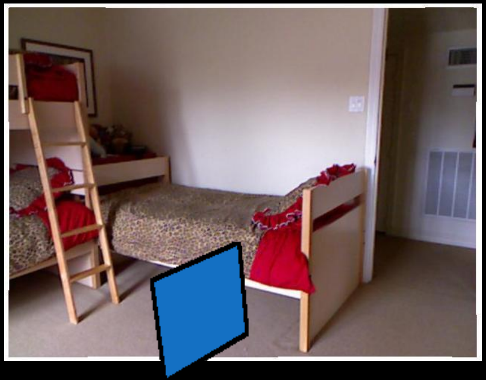
Formulation

$$\arg\max_{\mathbf{x}\in\{0,1\}^n} \mathbf{c}^T\mathbf{x} + \mathbf{x}^T\mathbf{H}\mathbf{x} \quad \text{s.t.} \quad \mathbf{A}\mathbf{x} \leq \mathbf{1}$$



Constraints

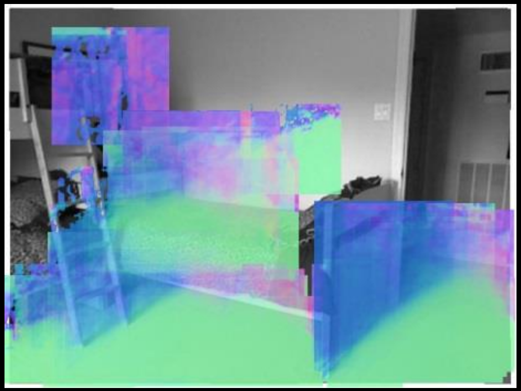
$$\arg\max_{\mathbf{x}\in\{0,1\}^n} \mathbf{c}^T\mathbf{x} + \mathbf{x}^T\mathbf{H}\mathbf{x} \quad \text{s.t.} \quad \mathbf{A}\mathbf{x} \leq \mathbf{1}$$



Unaries

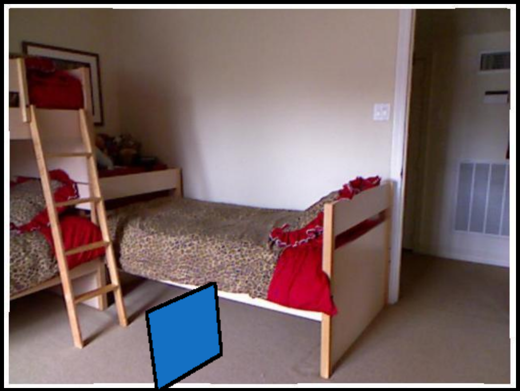
$$\arg\max_{\mathbf{x}\in\{0,1\}^n} \mathbf{c}^T\mathbf{x} + \mathbf{x}^T\mathbf{H}\mathbf{x} \quad \text{s.t.} \quad \mathbf{A}\mathbf{x} \leq \mathbf{1}$$

Unaries



Unary Evidence:
(1) 3DP
(2) Room Box Fitting

High c



Low c

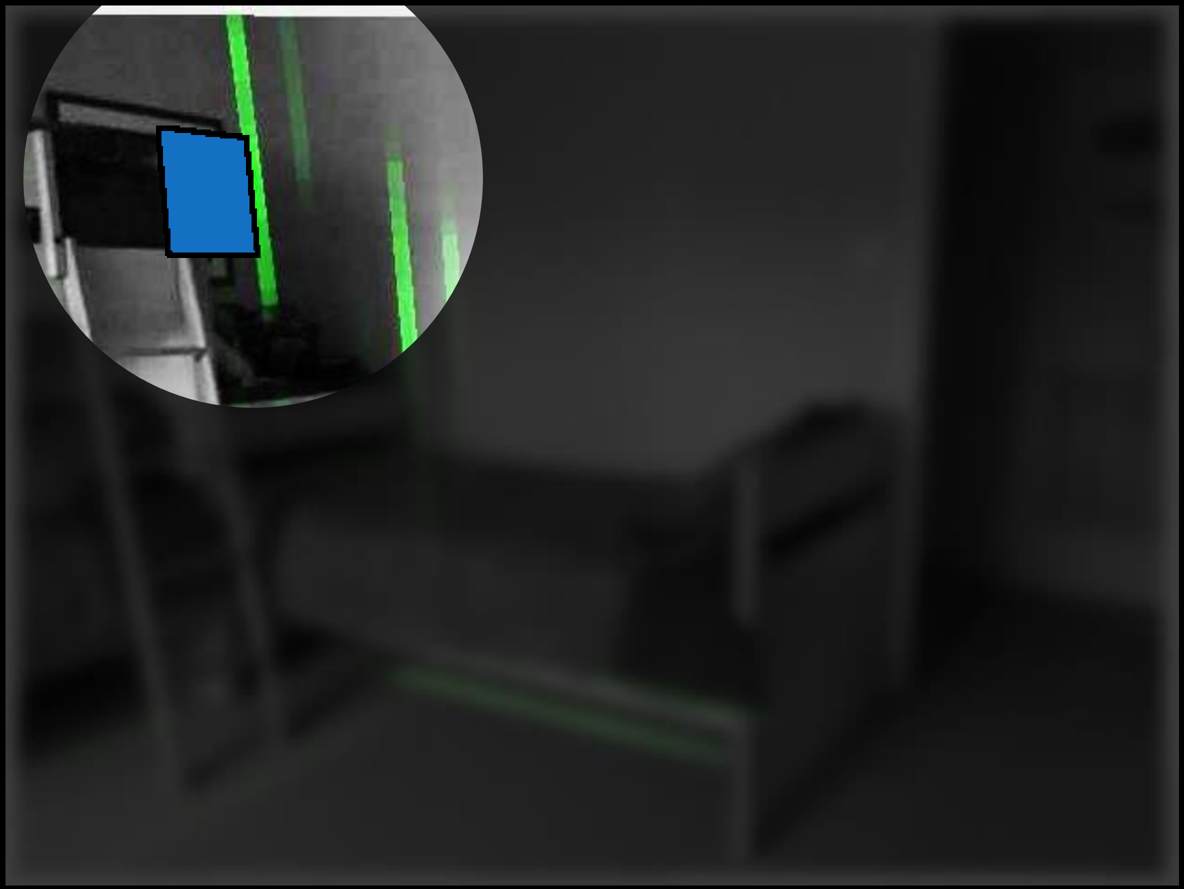
Binaries

$$\arg\max_{\mathbf{x}\in\{0,1\}^n} \mathbf{c}^T\mathbf{x} + \mathbf{x}^T\mathbf{H}\mathbf{x} \quad \text{s.t.} \quad \mathbf{A}\mathbf{x} \leq \mathbf{1}$$

Convex/Concave Constraints

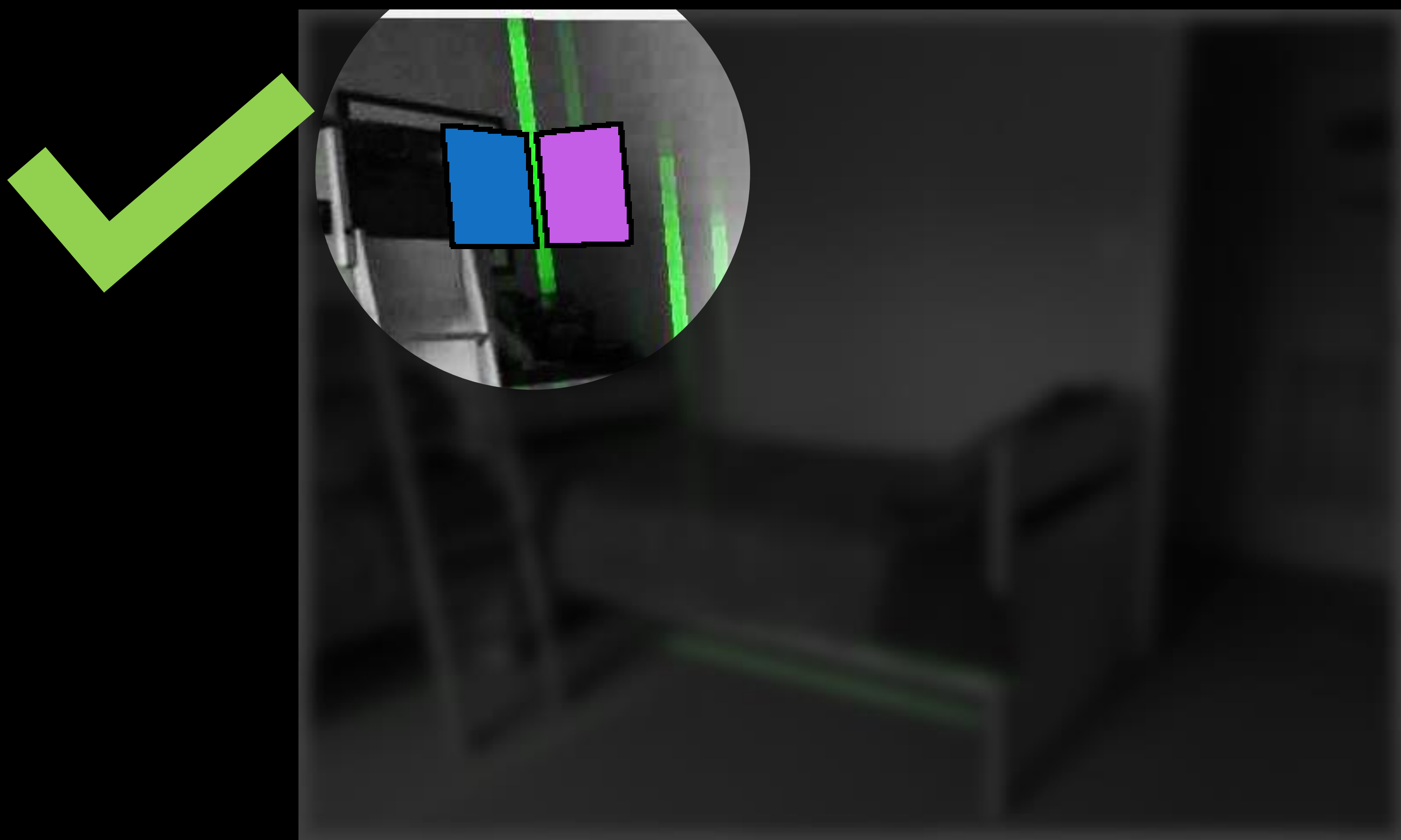
Detected Concave (-)

Convex/Concave Constraints



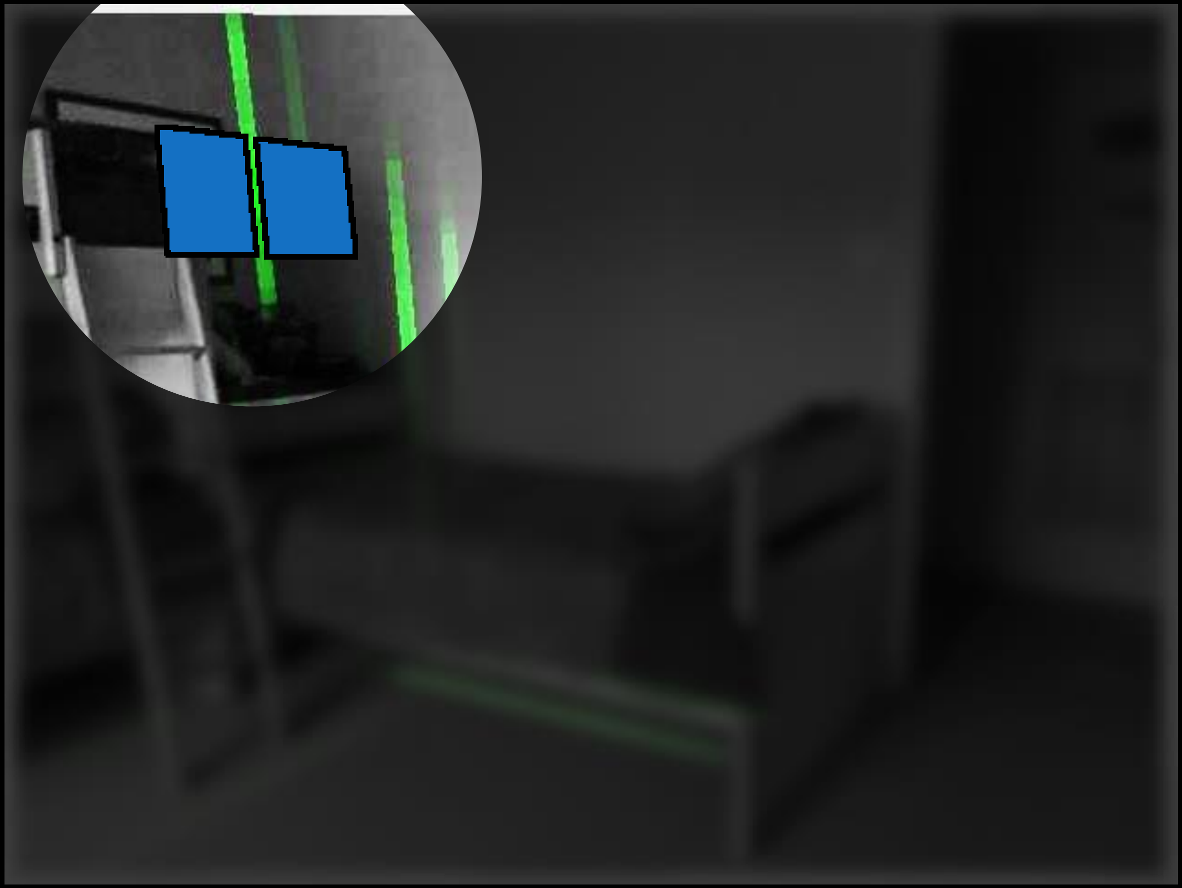
Detected Concave (-)

Convex/Concave Constraints



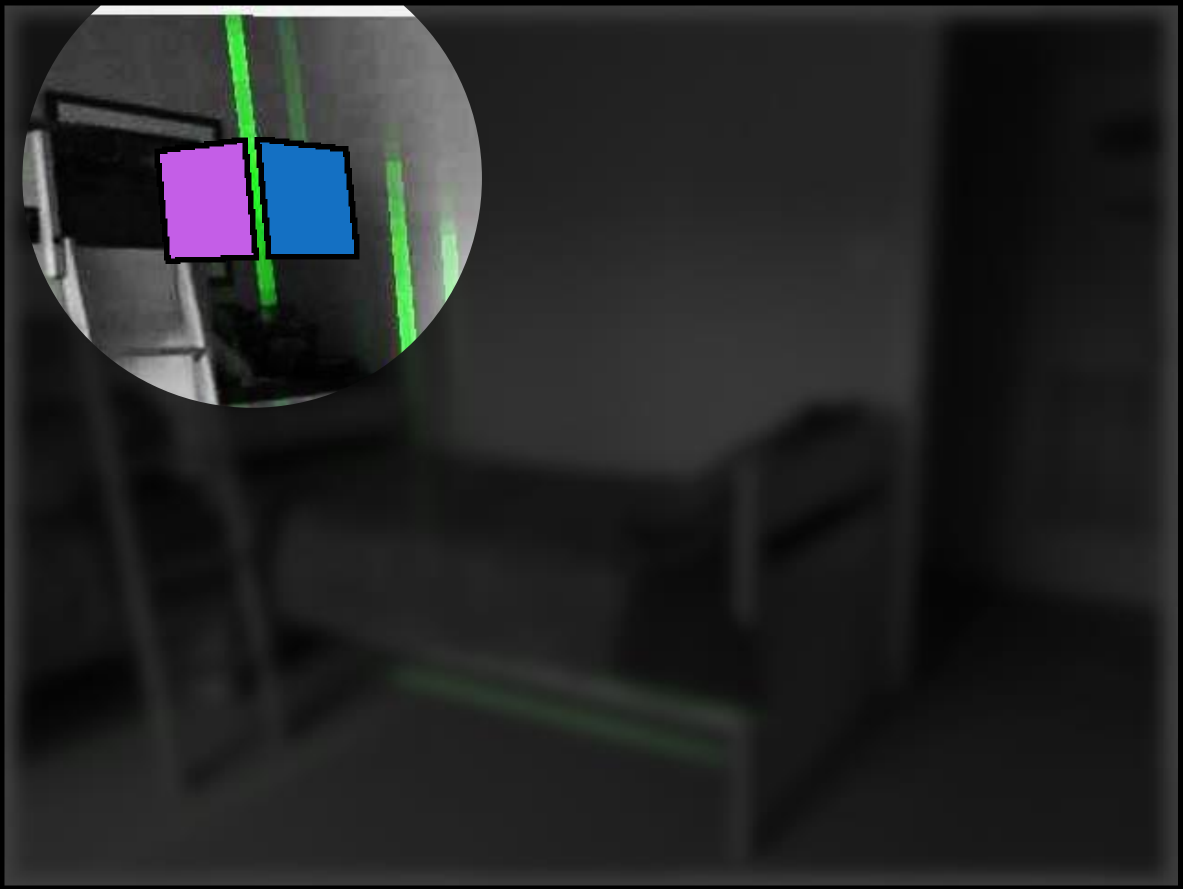
Detected Concave (-)

Convex/Concave Constraints



Detected Concave (-)

Convex/Concave Constraints

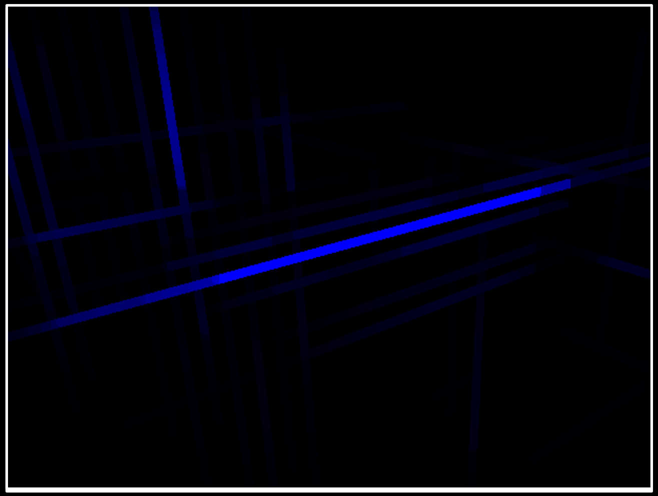


Detected Concave (-)

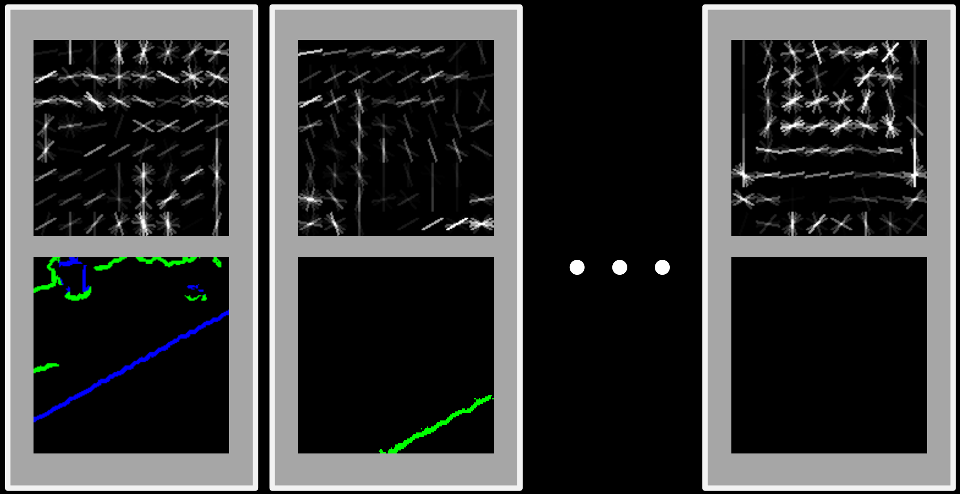
Detecting Convex/Concave

Use 3DP to Transfer Convex/Concave

Input

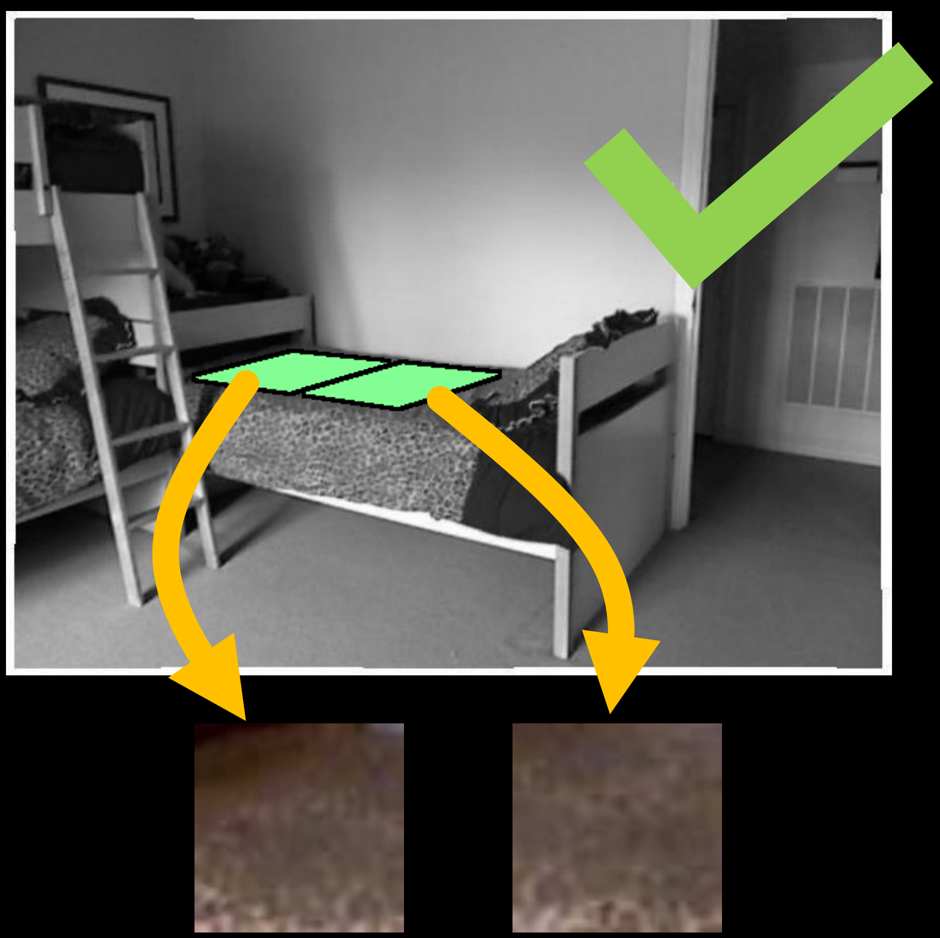


3D Primitive Bank



Ground-Truth Discontinuities similar to Gupta, Arbelaez, Malik, 2013
3DP from Fouhey, Gupta, Hebert, 2013

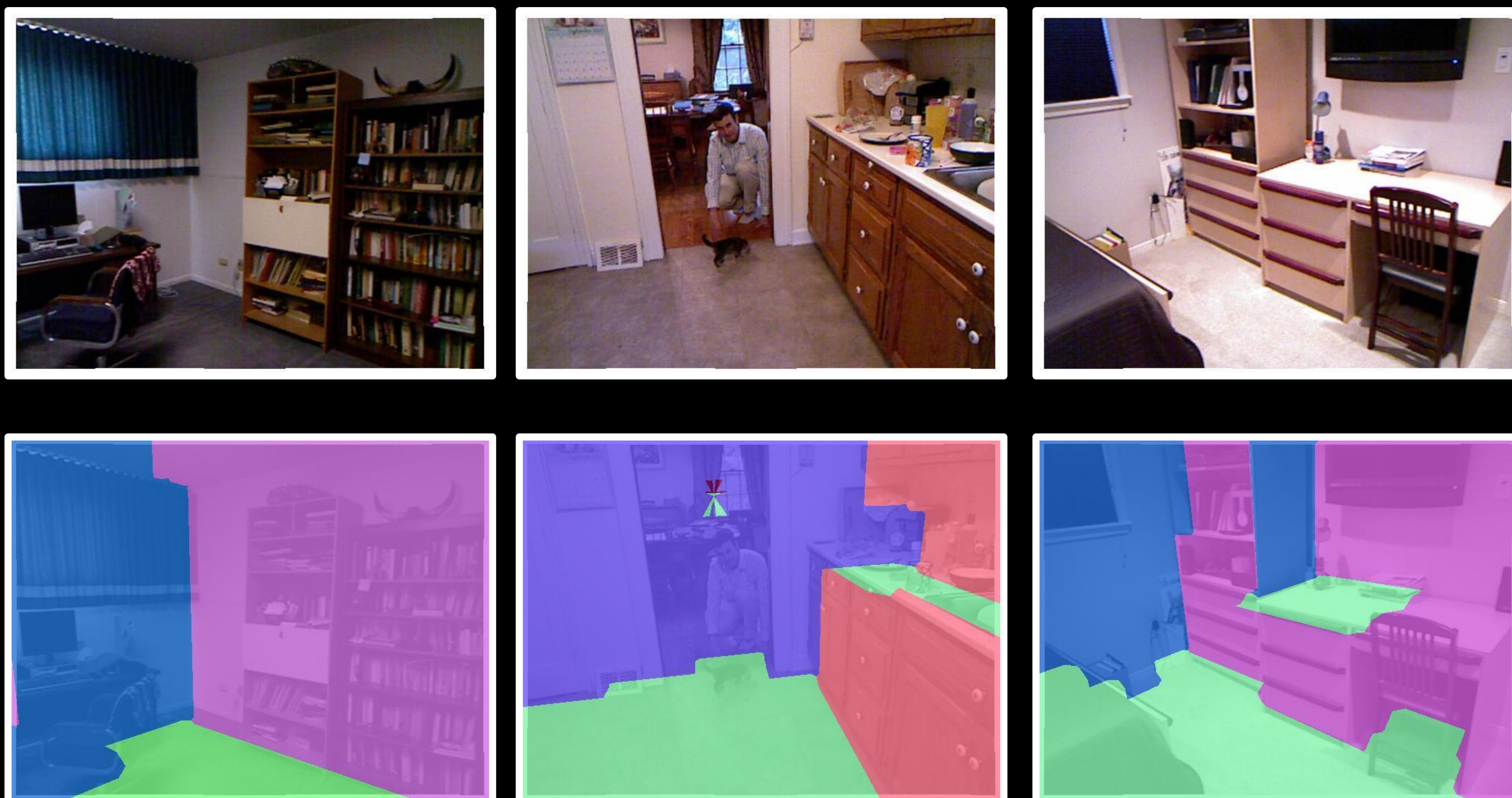
Smoothness



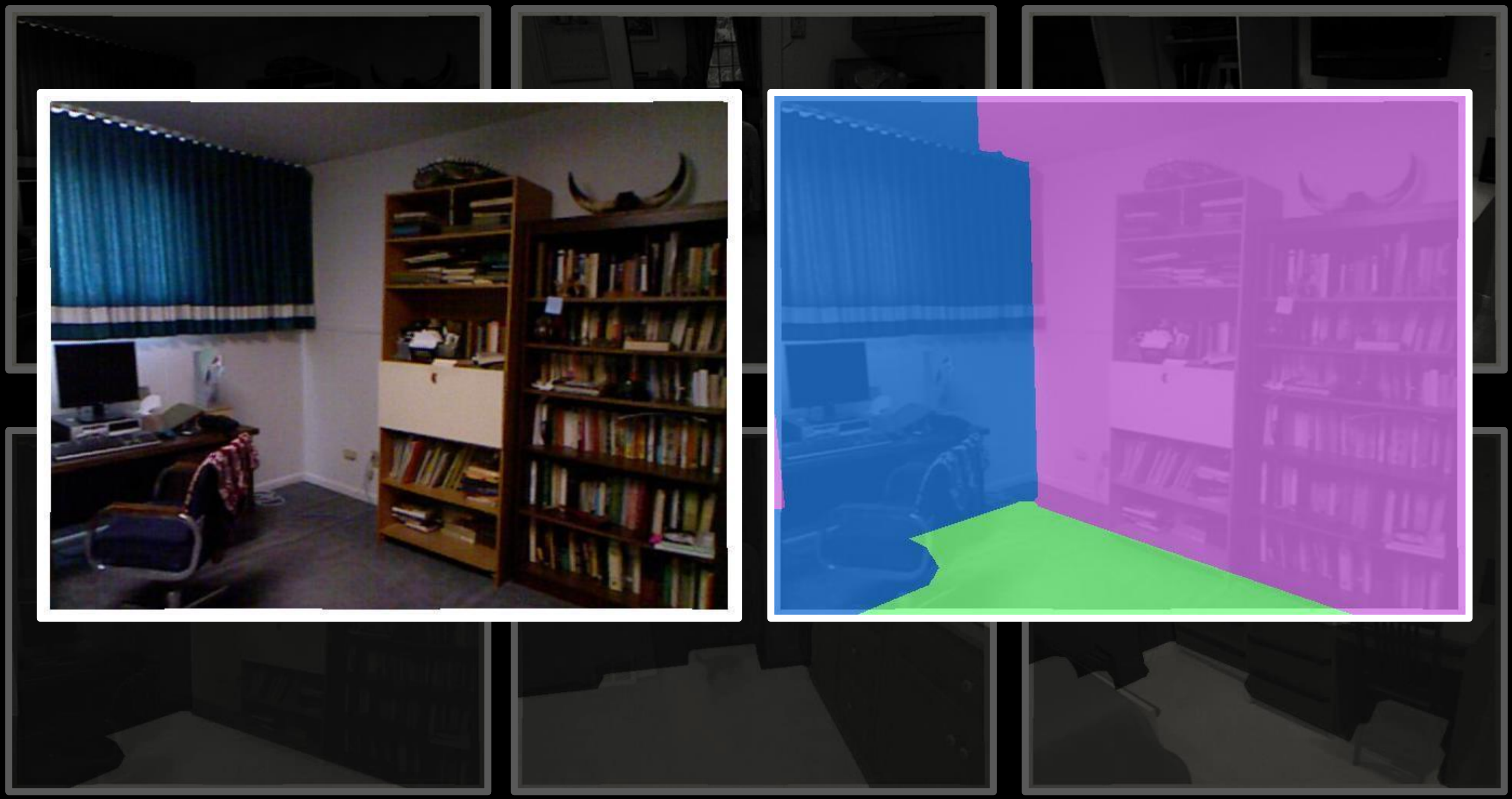
Solving the Model

$$\arg\max_{\mathbf{x}\in\{0,1\}^n} \mathbf{c}^T\mathbf{x} + \mathbf{x}^T\mathbf{H}\mathbf{x} \quad \text{s.t.} \quad \mathbf{A}\mathbf{x} \leq \mathbf{1}$$

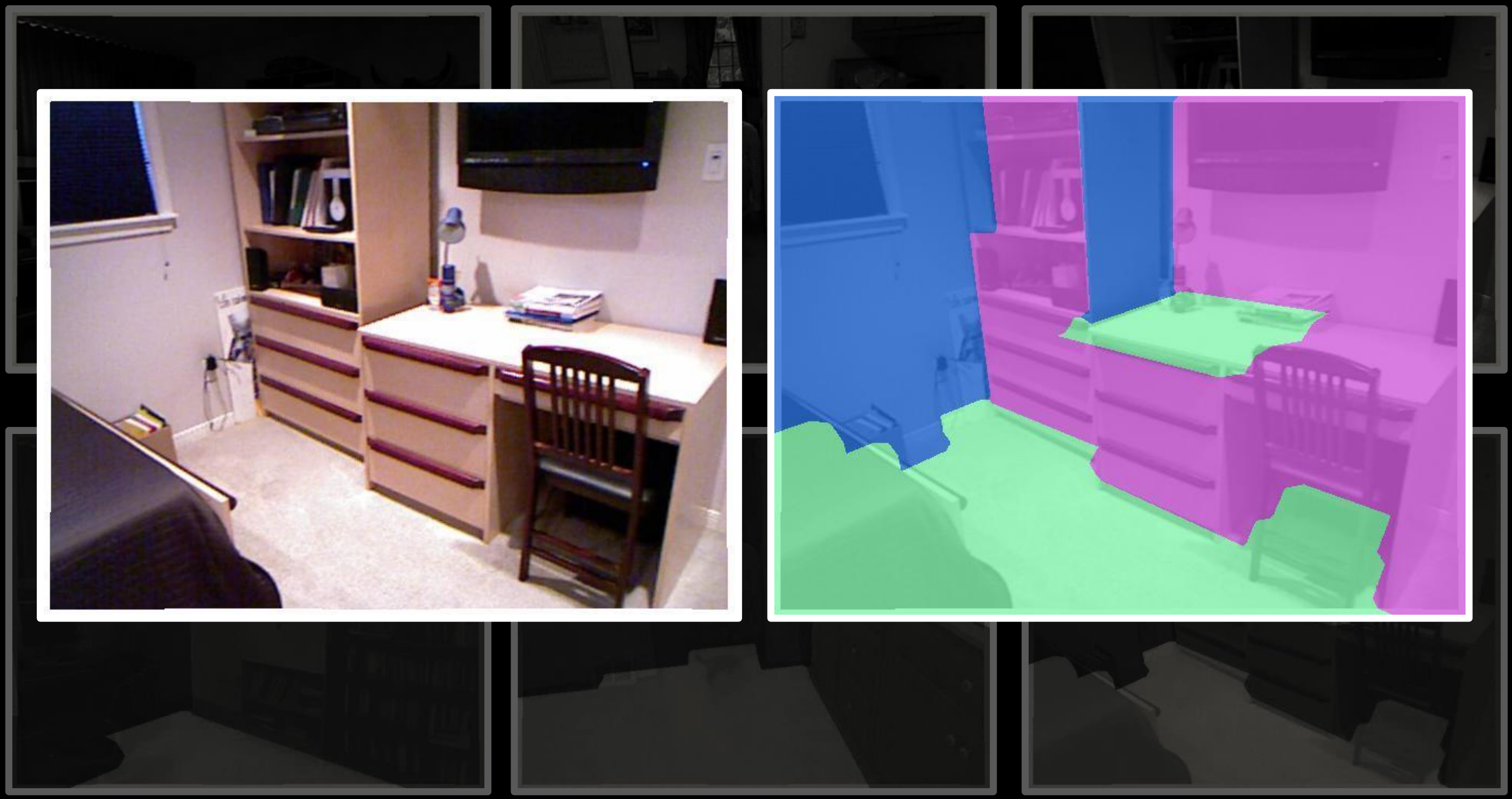
Qualitative Results



Qualitative Results



Qualitative Results



Results – Quantitative

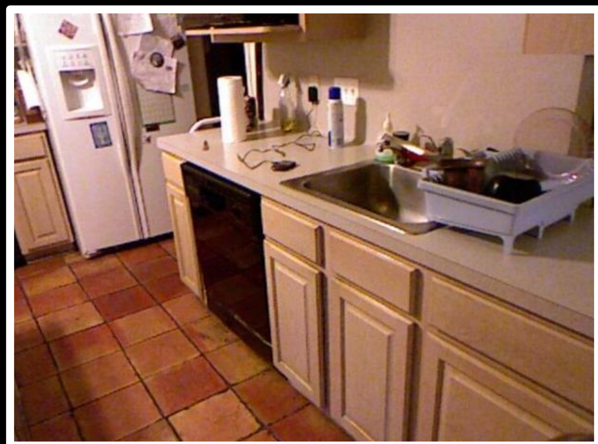
	Summary Stats (°) (Lower Better)		% Good Pixels (Higher Better)		
	Mean	Median	11.25°	22.5°	30°
Proposed	35.2	**17.9**	**40.5**	**54.1**	**58.9**
3DP	36.3	19.2	39.2	52.9	57.8
Ladicky '14	33.5	23.1	27.7	49.0	58.7

Fouhey et al. ICCV '13; Ladicky et al. ECCV '14

CONCLUSIONS & FUTURE WORK

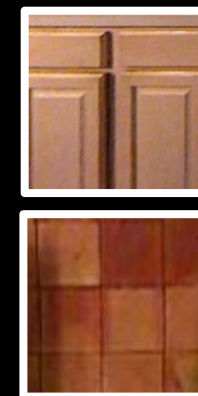
Today

Image (3D Structure x Style)

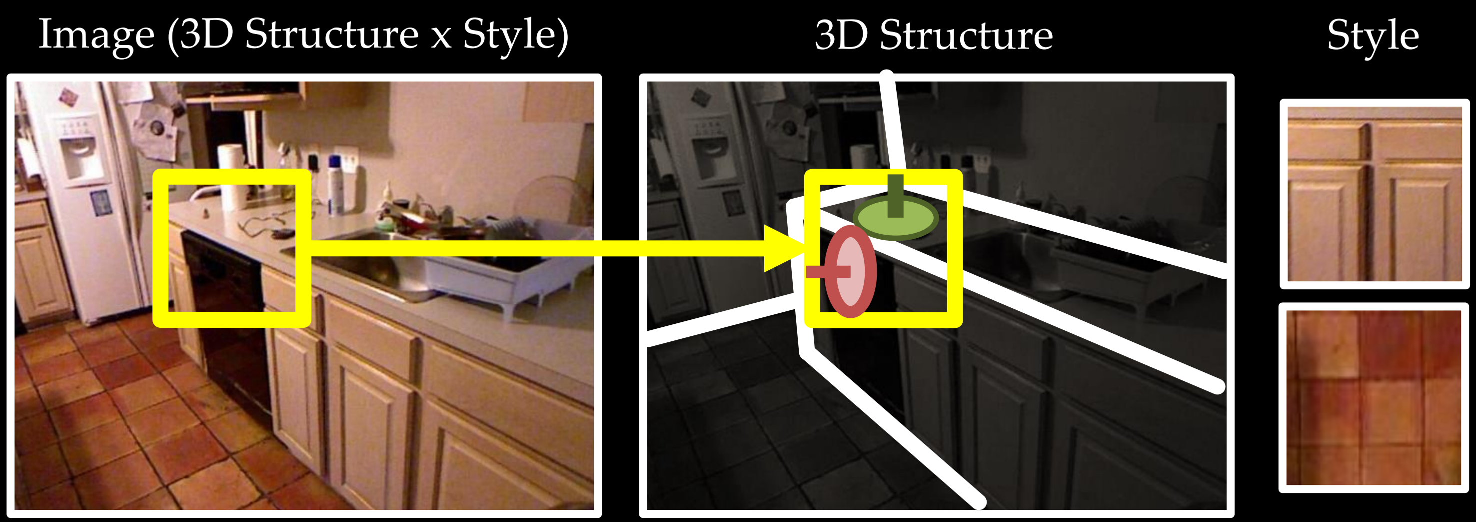


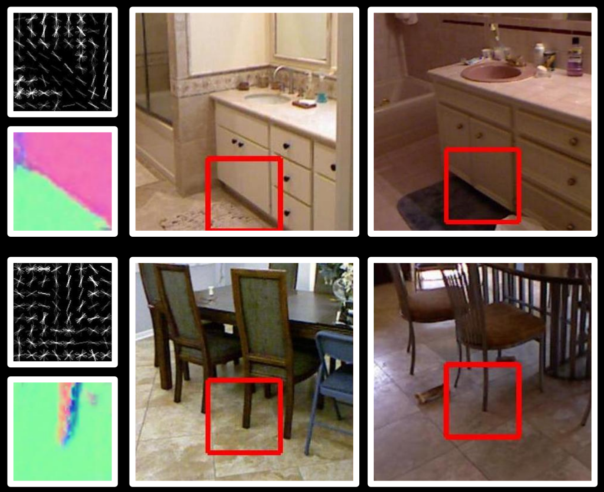
3D Structure

Style



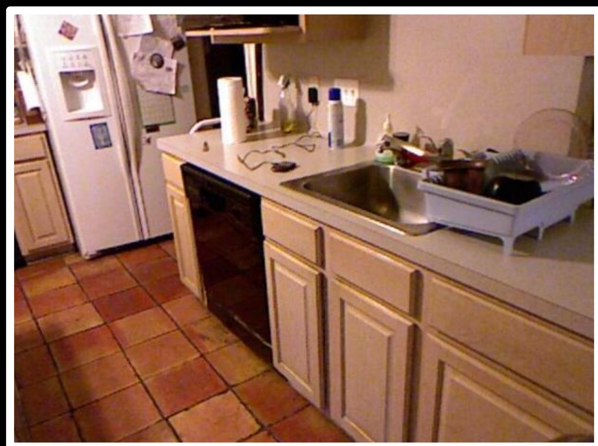
Today



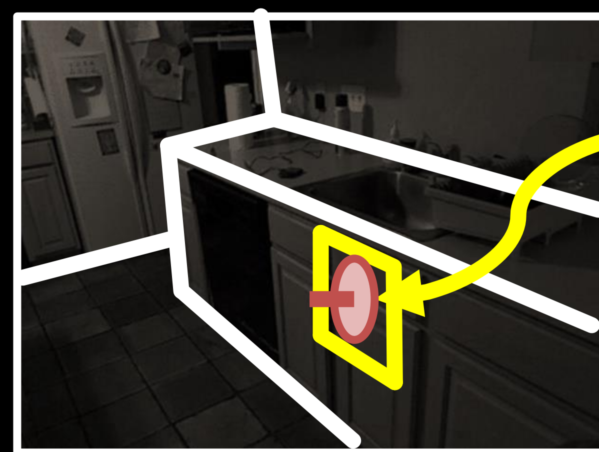


Today

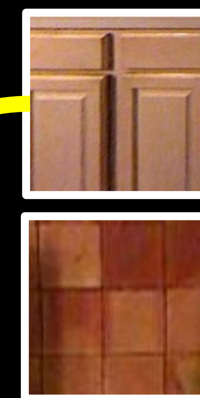
Image (3D Structure x Style)

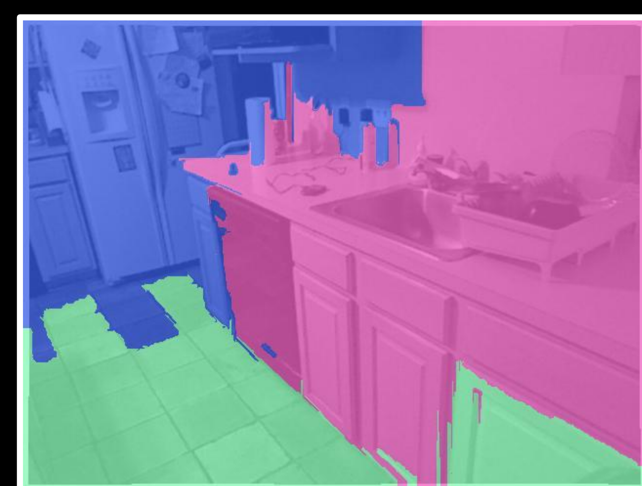


3D Structure



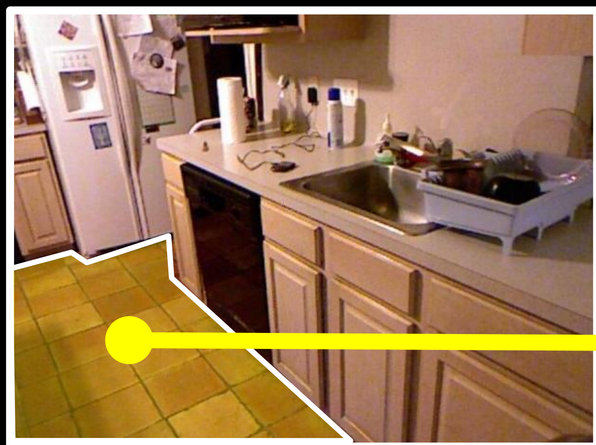
Style





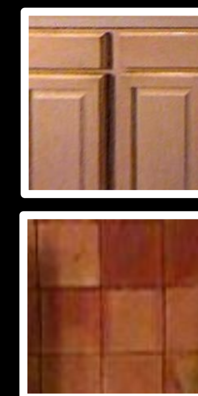
Today

Image (3D Structure x Style)



3D Structure

Style

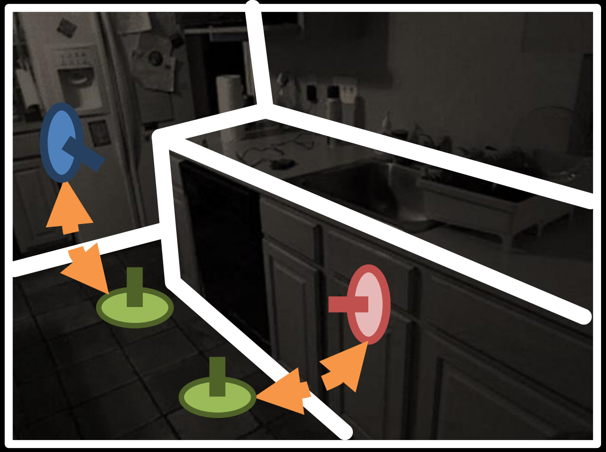


Planar	✓
Non-Planar	✗
Cylindrical	✗
Rough Surf	✗
Pnt/L Contact	✓
Mult. Contact	✓
Empty	✗
Mult. Pieces	✓
Holes	✓
Thin	✓
Mirror Sym.	✓
Cubic Aspect	✗

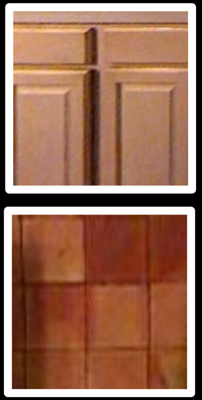
Today

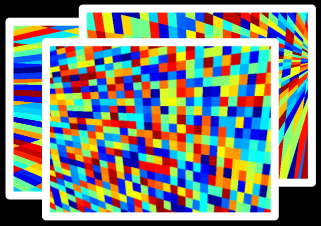
Image (3D Structure x Style)

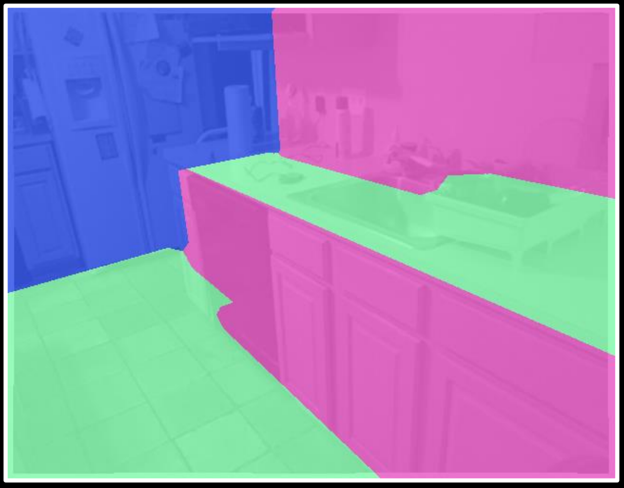
3D Structure



Style



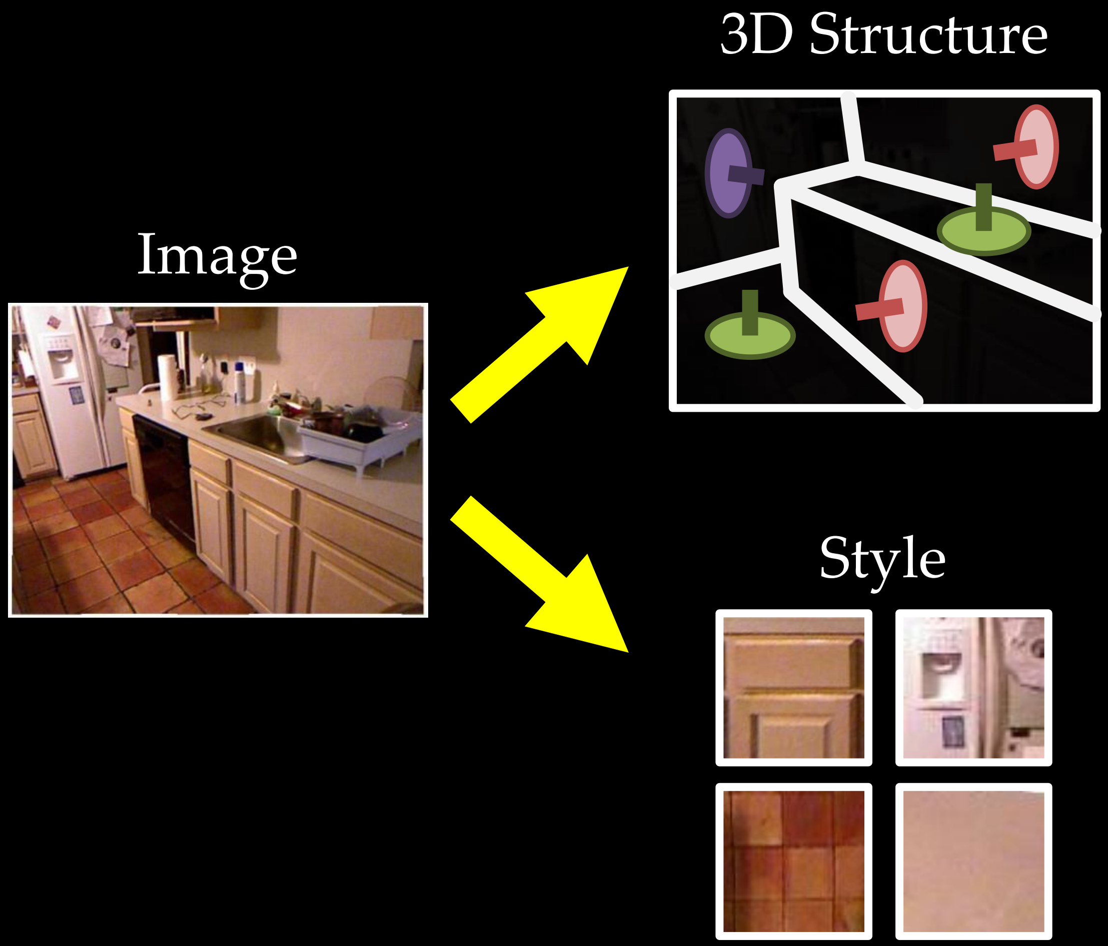




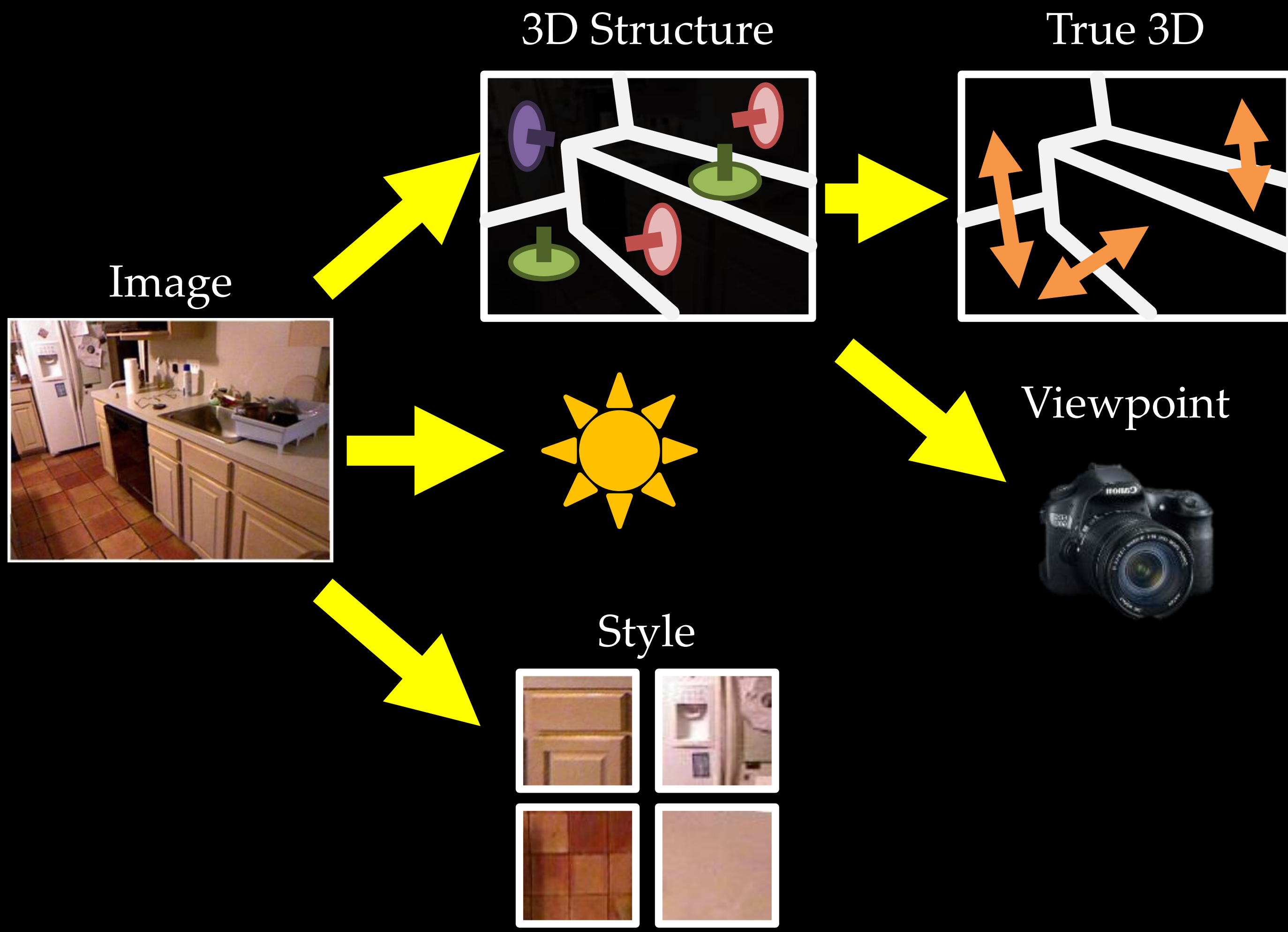
$$\arg\max_{\mathbf{x}\in\{0,1\}^n} \mathbf{c}^T\mathbf{x} + \mathbf{x}^T\mathbf{H}\mathbf{x}$$
$$\text{s.t.} \quad \mathbf{A}\mathbf{x} \leq \mathbf{1}$$

Future Work

Further Factorization

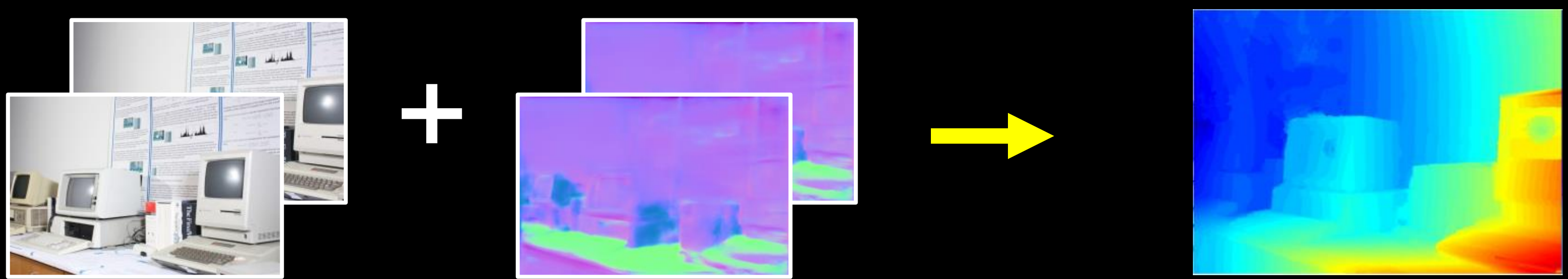


Further Factorization



Reuniting 3Ds (Multiview)

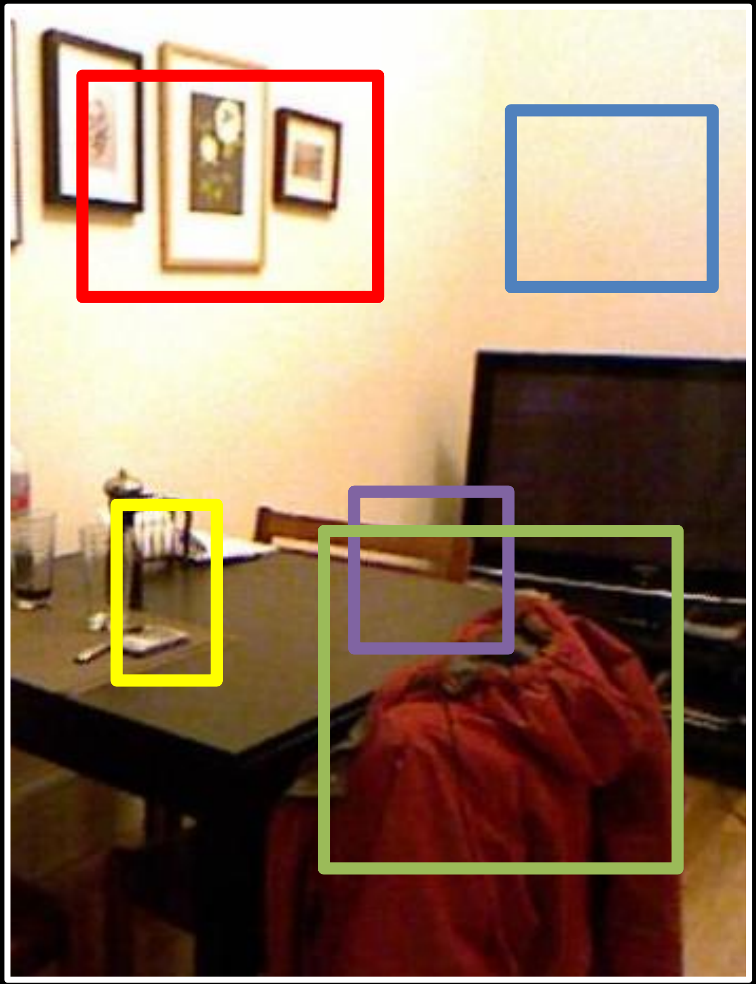
Monocular and multi-view cues



Supervised and unsupervised models

E.g., Concha et al. Autonomous Robots ‘15, Hadfield et al. ICCV ‘15, Hane et al. CVPR ‘15

Reuniting 3Ds (Single View)



Texture

Top-down

Occlusion

Shading

Semantics

Thank you