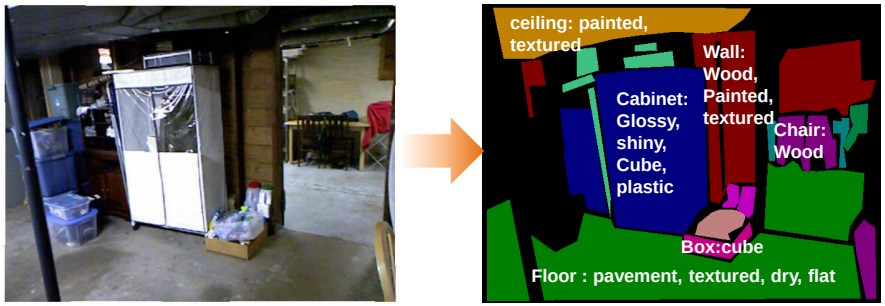
ImageSpirit: Verbal Guide Image Parsing

Shuai Zheng

Joint work with Ming-Ming Cheng, Wen-Yan Lin, Vibhav Vineet, Paul Sturgess, Nigel Crook, Niloy Mitra, Carsten Rother, and Philip Torr

Torr Vision Group, University of Oxford

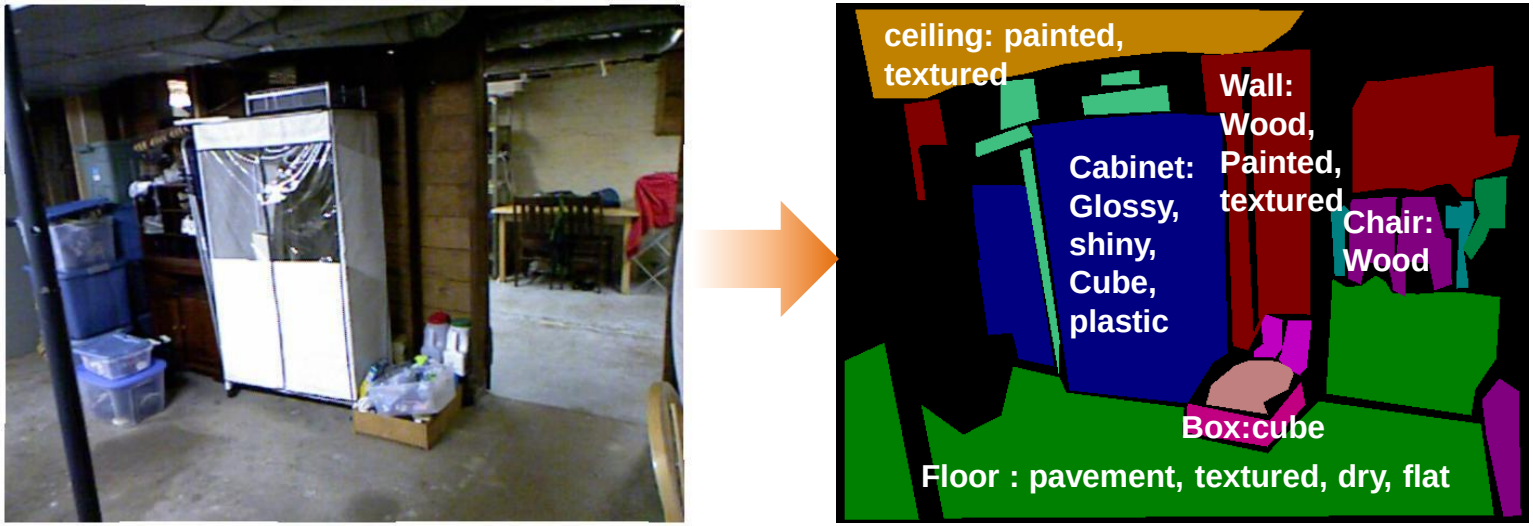
Content



Dense Semantic Image Segmentation with Objects and Attributes

Verbal Guided Image Parsing

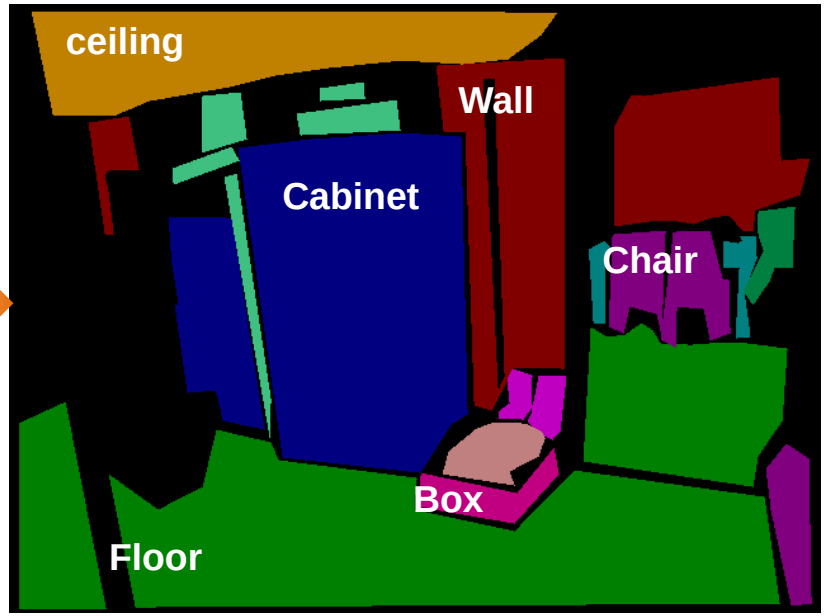
Dense Semantic Image Segmentation with Objects and Attributes



Dense Semantic Image Segmentation with Objects and Attributes,, IEEE Computer Vision and Pattern Recognition, 2014.

Goal of Traditional Image Parsing

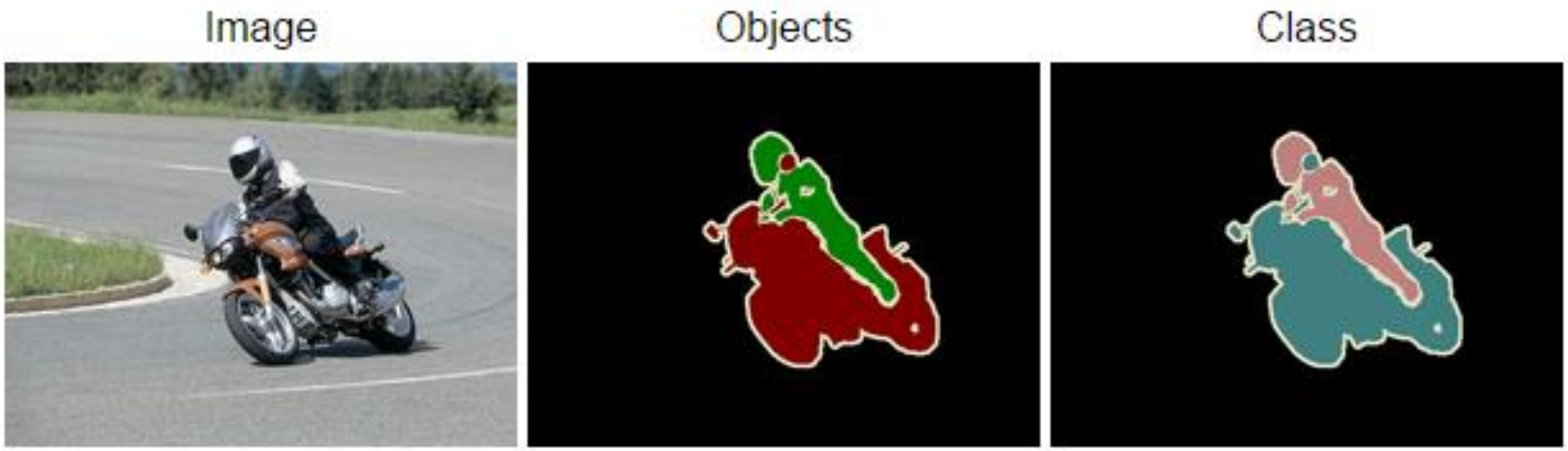
- Object class segmentation
  - Assigning an object class label to each pixel



Source: NYU V2 Image parsing ground truth.

Goal of Traditional Image Parsing

- Image Parsing (Semantic Image Segmentation)
  - Generating pixel-wise segmentations giving the class of the object visible at each pixel, or "background" otherwise.



PASCAL VOC Semantic Image Segmentation Competition

Visual Attributes

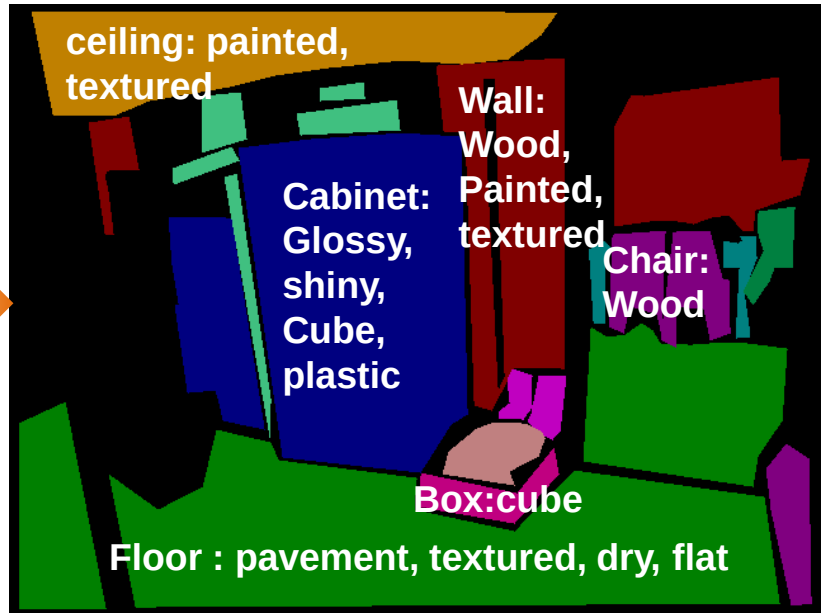
- ***Visual attributes*** are properties observable in images that have human-designated names, such as ‘Orange’, ‘striped’, or ‘Furry’.

4-Legged	White	Male
Orange	Symmetric	Asian
Striped	Ionic columns	Beard
Furry	Classical	Smiling

Source: Learning visual attributes. NIPS 2007. & Relative Attributes. ICCV 2011.

Our Goal

- Segmentation with Objects and Attributes
  - Assigning an object class and a set of attribute labels to each pixel



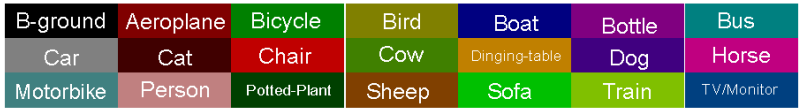
Source: Attribute-augmented NYU V2 Image parsing ground truth.

Labeling Problem

The goal of labeling problem is to find a set of labelling that maximizes the conditional probability, or minimize the energy function.

$$E(\boldsymbol{z}) = \overbrace{\sum_{i \in I} \psi_i(z_i)}^{Unary} + \overbrace{\sum_{i \neq j \in I} \psi_{ij}(z_i, z_j)}^{Pairwise}$$

$$Z^* = \arg\max_{z} P(z|I) = \arg\min_{z} E(z|I)$$

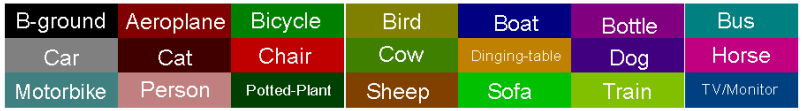


Labeling Problem

The goal of labeling problem is to find a set of labelling that maximizes the conditional probability, or minimize the energy function.

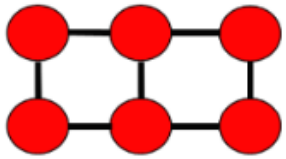
$$\boldsymbol{E}(\boldsymbol{z}) = \overbrace{\sum_{i \in I} \psi_i(z_i)}^{Unary} + \overbrace{\sum_{i \neq j \in I} \psi_{ij}(z_i, z_j)}^{Pairwise}$$

$$Z^* = \arg\max_z P(z|I) = \arg\min_z E(z|I)$$

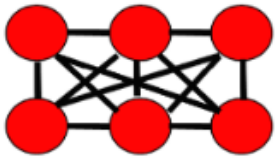


CRF Inference

Move-Making	Message Passing
• Boykov et al. Fast approximate energy minimization via graph cuts. PAMI 2001.	• Murphy. Loopy belief propagation: An empirical study, UAI, 1999. • T. Minka. Expectation propagation for approximate bayesian inference. UAI, 2001.
• Ladicky et al. Associative Hierarchical CRFs for Object Class Image Segmentation. ICCV 2009.	• Krähenbühl. Efficient Inference in Fully Connected CRFs with Gaussian Edge Potentials. NIPS 2011.



Grid CRF



Fully-connected CRF

Dense CRF

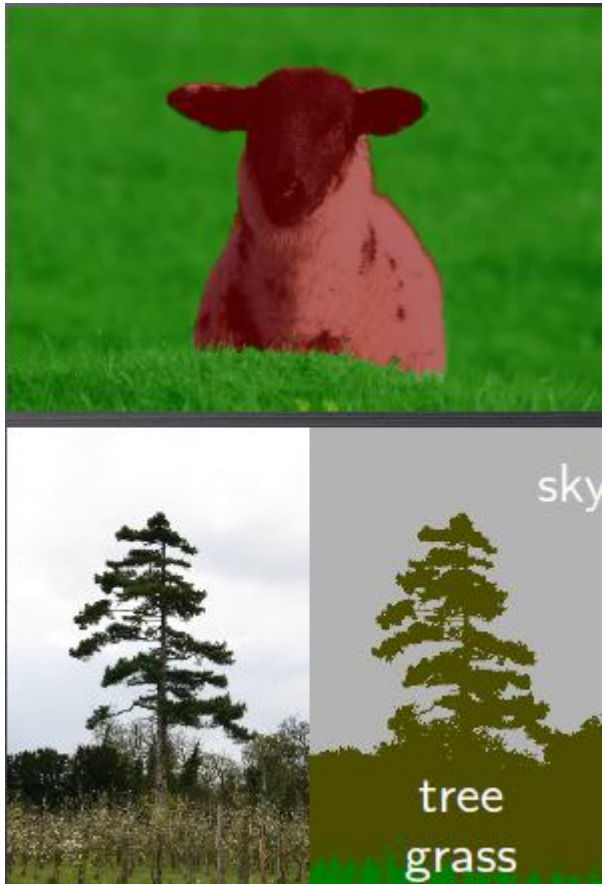
Dense Random Fields

Pros:

- Long range interactions
- Probabilistic interpretation
- Higher order potentials
- Parameter Learning

Challenges:

- Verbal Large model
  - 50,000+ variables
  - Billions of pairwise terms
- Traditional Inference very slow
  - MCMC partially converges in 36hrs
  - GraphCuts + alpha expansion does not converge in 3 days.



Source: Efficient Inference in Fully Connected CRFs with Gaussian Edge Potentials. NIPS 2011.

Dense CRF

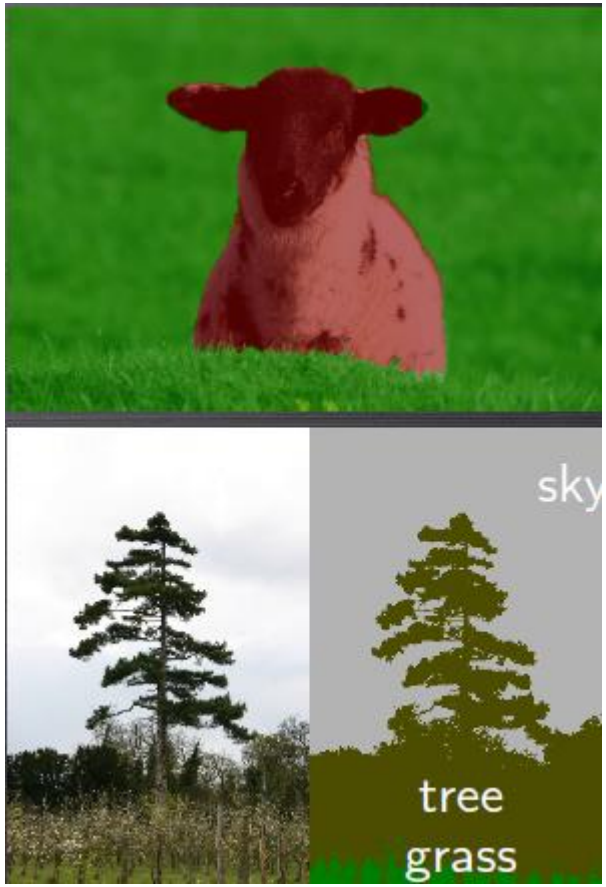
Efficient Inference in Dense Random Fields

Efficient Inference

- 0.2s per frame

Pairwise term

- Linear combination of Gaussians



Source: Efficient Inference in Fully Connected CRFs with Gaussian Edge Potentials. NIPS 2011.

Dense CRF

$$\boldsymbol{E}(\boldsymbol{z}) = \overbrace{\sum_{i \in I} \psi_i(z_i)}^{Unary} + \overbrace{\sum_{i \neq j \in I} \psi_{ij}(z_i, z_j)}^{Pairwise}$$

Pairwise term:

$$\psi_{ij}(z_i, z_j) = \overbrace{\mu(z_i, z_j)}^{Label\ compatibility} \left(w_1 \overbrace{\exp\left(-\frac{|p_i - p_j|}{2\theta_\alpha^2} - \frac{|I_i - I_j|}{2\theta_\beta^2}\right)}^{color-sensitive\ model} + w_2 \underbrace{\exp\left(-\frac{|p_i - p_j|}{2\theta_\gamma^2}\right)}_{Local\ Smooth\ model}\right)$$

Dense CRF

- Mean-field Inference

$Q_i \leftarrow \frac{1}{Z_i} \exp\left(U_i(l)\right)$ for all i ▷ Initialization O(N)

while not converged **do**

$\tilde{Q}_i^{(m)}(l) \leftarrow \sum_{j \neq i} k^{(m)}(\mathbf{f}_i, \mathbf{f}_j) Q_j(l)$ for all m

▷ Message Passing O(N^2)

$\check{Q}_i(l) \leftarrow \sum_m w^{(m)} \tilde{Q}_i^{(m)}(l)$

▷ Weighting Filter Outputs

$\hat{Q}_i(l) \leftarrow \sum_{l' \in \mathcal{L}} \mu(l, l') \check{Q}_i(l)$ O(N)

▷ Compatibility Transform

$\check{Q}_i(l) \leftarrow U_i(l) - \hat{Q}_i(l)$

▷ Adding Unary Potentials O(N)

$Q_i \leftarrow \frac{1}{Z_i} \exp\left(\check{Q}_i(l)\right)$

▷ Normalizing O(N)

end while

Source: Efficient Inference in Fully Connected CRFs with Gaussian Edge Potentials. NIPS 2011.

Dense CRF

- Efficient Mean-field Inference

$Q_i \leftarrow \frac{1}{Z_i} \exp\left(U_i(l)\right)$ for all i ▷ Initialization O(N)

while not converged **do**

High Dimensional Filter

▷ Message Passing O(N)

$\check{Q}_i(l) \leftarrow \sum_m w^{(m)} \tilde{Q}_i^{(m)}(l)$

▷ Weighting Filter Outputs

$\hat{Q}_i(l) \leftarrow \sum_{l' \in \mathcal{L}} \mu(l, l') \check{Q}_i(l)$ O(N)

▷ Compatibility Transform

$\check{Q}_i(l) \leftarrow U_i(l) - \hat{Q}_i(l)$

▷ Adding Unary Potentials O(N)

$Q_i \leftarrow \frac{1}{Z_i} \exp\left(\check{Q}_i(l)\right)$

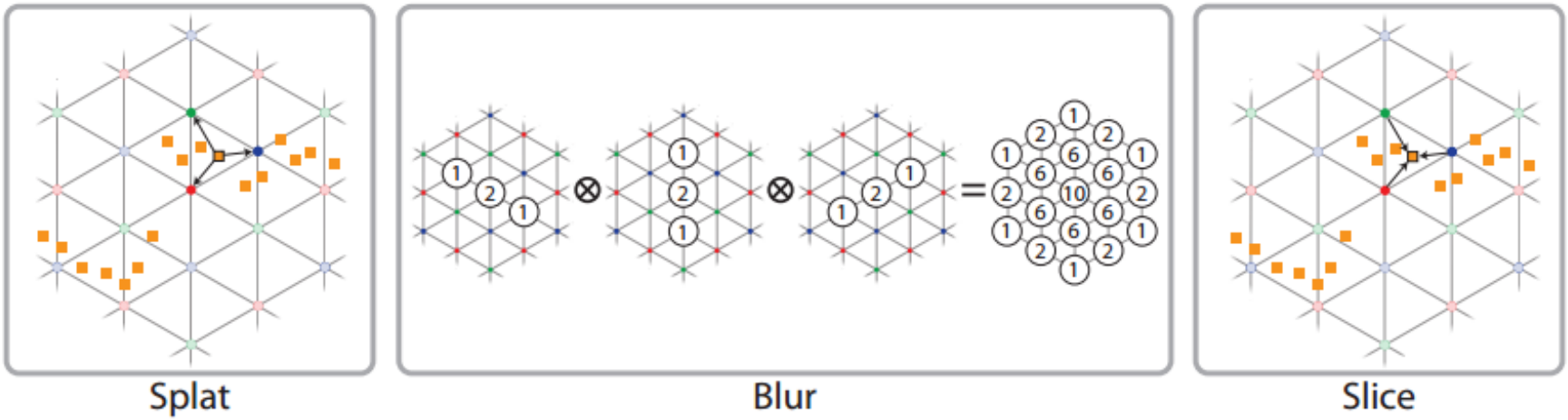
▷ Normalizing O(N)

end while

Source: Fast High-Dimensional Filtering Using the Permutohedral Lattice. Eurographics, 2010.

Permutohedral lattice

- Efficient High-Dimensional filtering with Permuohedral lattice



Source: Fast High-Dimensional Filtering Using the Permutohedral Lattice. Eurographics, 2010.

Multi-Label Factorial Dense CRF

$$E(\mathbf{z}) = \overbrace{\sum_{i \in I} \psi_i(z_i)}^{Unary} + \overbrace{\sum_{i \neq j \in I} \psi_{ij}(z_i, z_j)}^{Pairwise}$$

$$\psi_i(z_i) = \overbrace{\psi_i^O(x_i)}^{Object} + \overbrace{\sum_a \psi_{i,a}^A(y_{i,a})}^{Attributes} + \overbrace{\sum_{o,a} \psi_{i,o,a}^{OA}(x_i, y_{i,a})}^{Joint\ Object-Attribute} + \overbrace{\sum_{a \neq a'} \psi_{i,a,a'}^A(y_{i,a}, y_{i,a'})}^{Joint\ Attribute-Attribute}$$

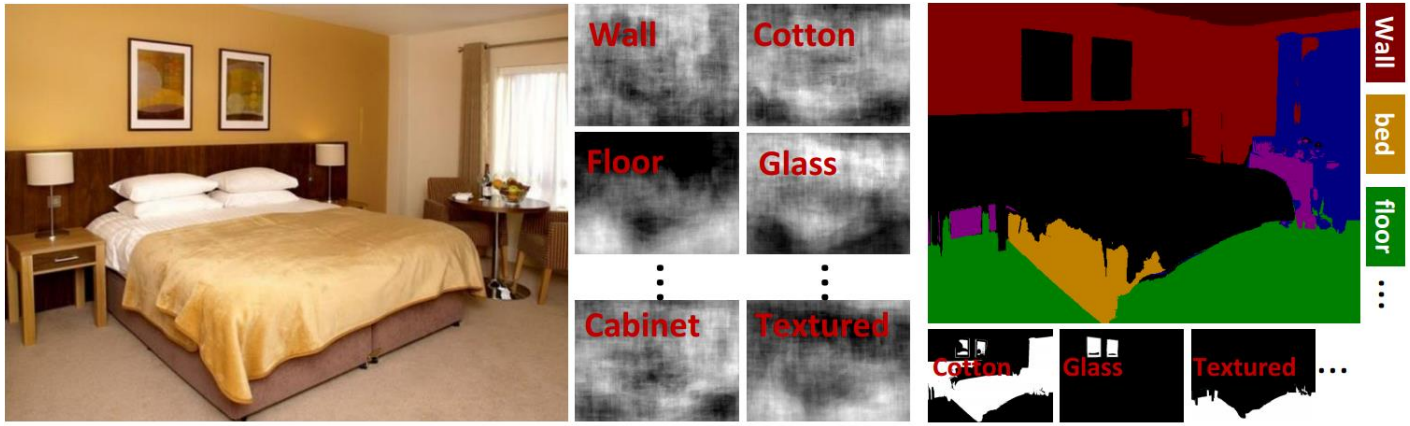
Object classifiers: table, chair, etc.

Attributes classifiers: wood, plastic, red, etc.

Object and attributes correlation.

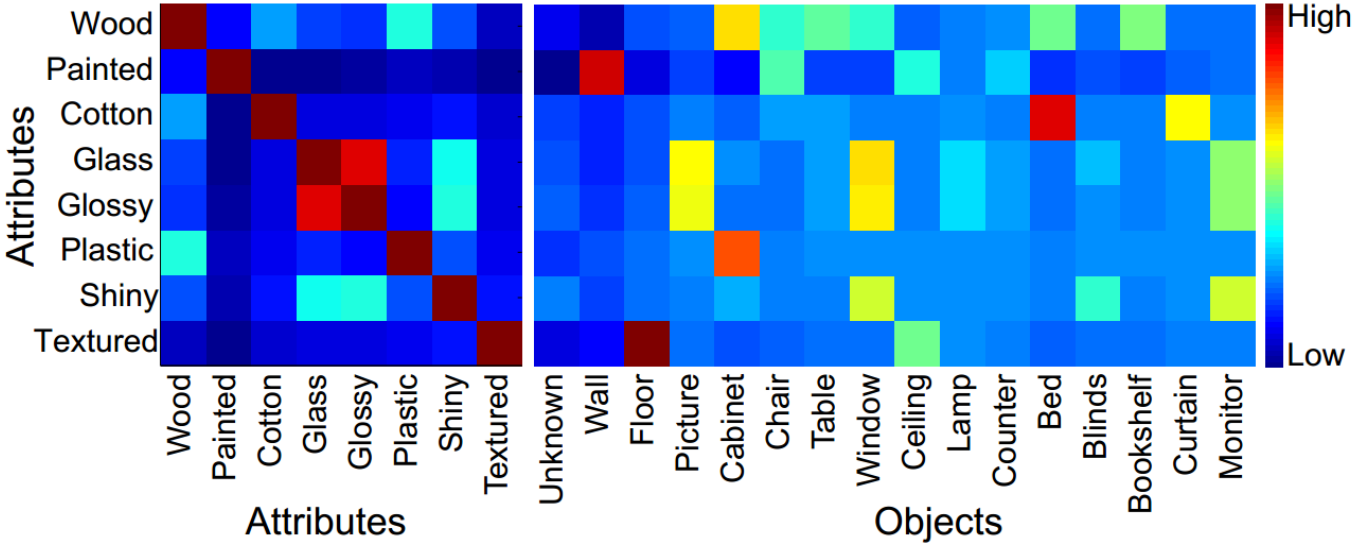
Correlation between attributes.

Joint Inference Results



(a) Inputs: an image and learned weak hypothesis [Shotton et al. 2009]

(b) Automatic scene parsing results

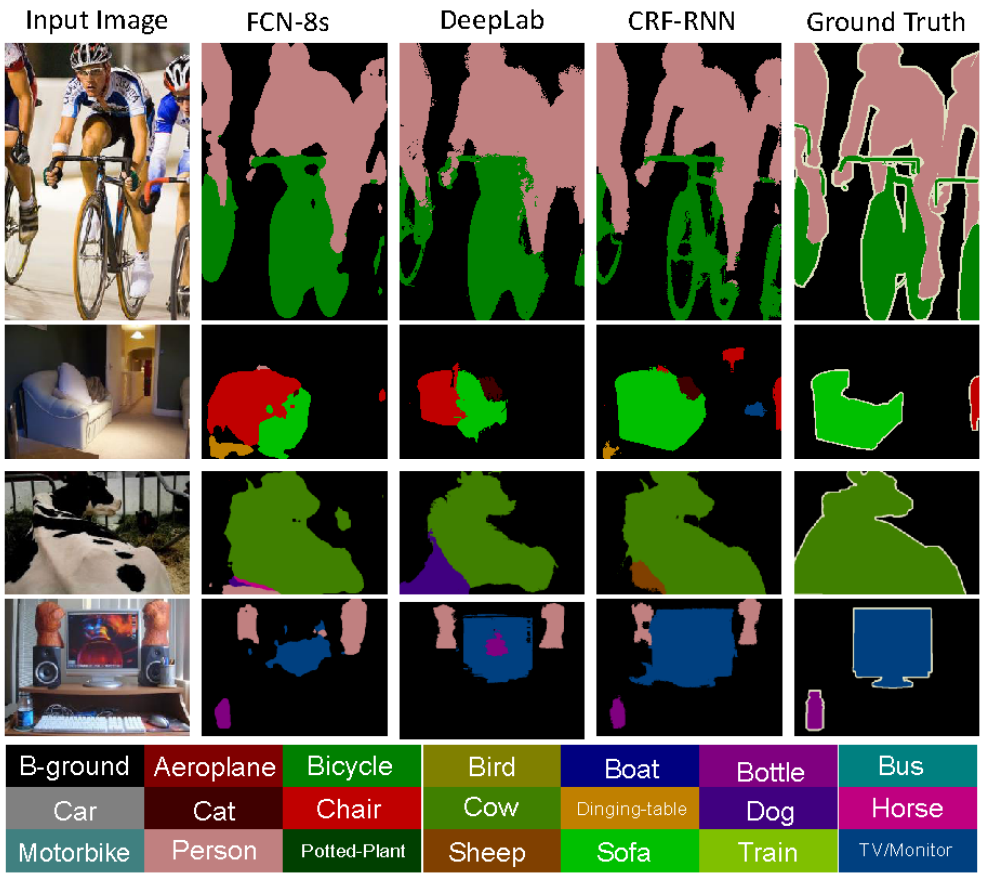


Verbal Guided Image Parsing

ImageSpirit: Verbal Guided Image Parsing, ACM Transactions on Graphics, 2014.

Motivations

- when image parsing is still far from perfect, how can we empower users to editing images? (



S. Zheng et al. Conditional Random Fields as Recurrent Neural Networks. arXiv:1502.03240. 2015
http://www.robots.ox.ac.uk/~szheng/CRFasRNN.html

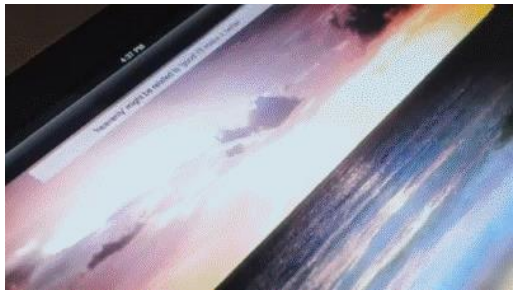
Motivation

- Concurrent work: PixelTone
  - Sketch contour + speech commands, etc.

Source: PixelTone: A Multimodal Interface for Image Editing. ACM CHI. 2013.

Motivation

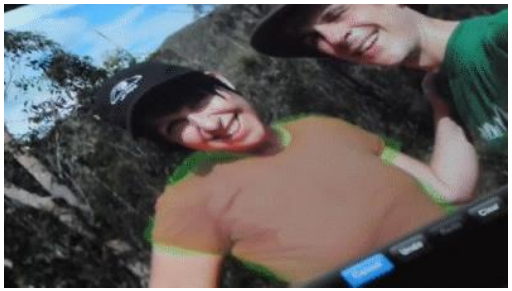
- Concurrent work: PixelTone
  - Sketch contour + speech commands, etc.



Source: PixelTone: A Multimodal Interface for Image Editing. ACM CHI. 2013.

Motivation

- Concurrent work: PixelTone
  - Sketch contour + speech commands, etc.



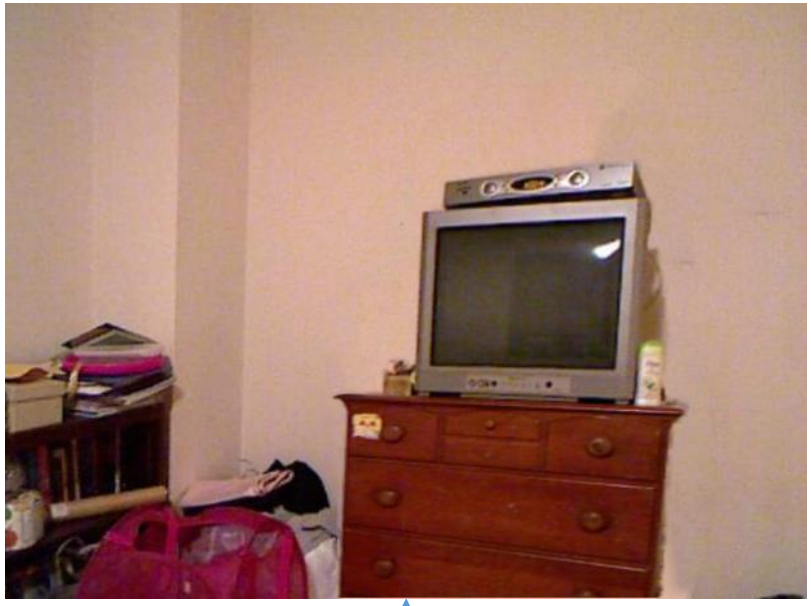
Source: PixelTone: A Multimodal Interface for Image Editing. ACM CHI. 2013.

Motivation

- Concurrent work: PixelTone
  - Sketch contour + speech commands, etc.

Source: PixelTone: A Multimodal Interface for Image Editing. ACM CHI. 2013.

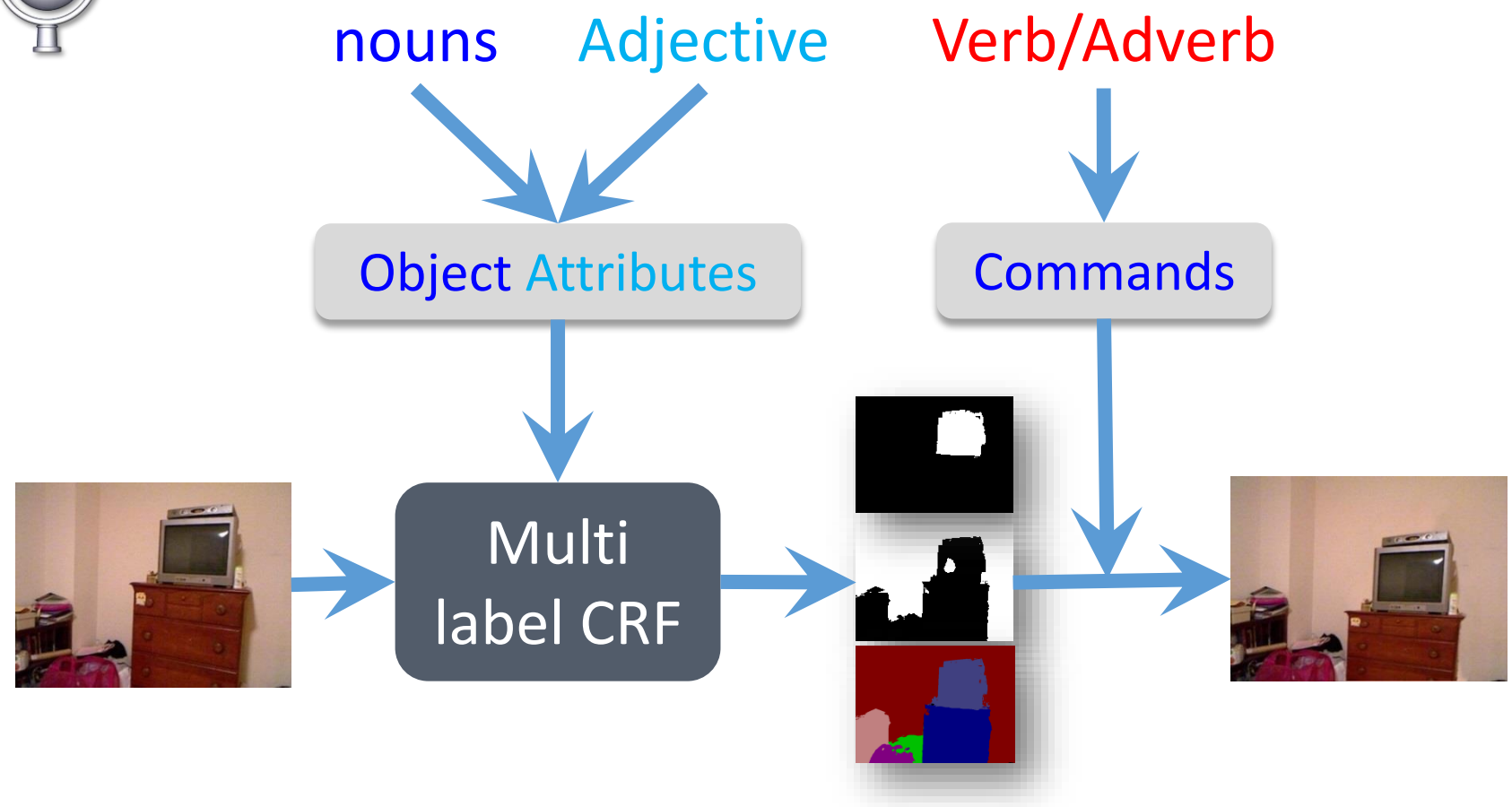
Verbal Guided Image Parsing



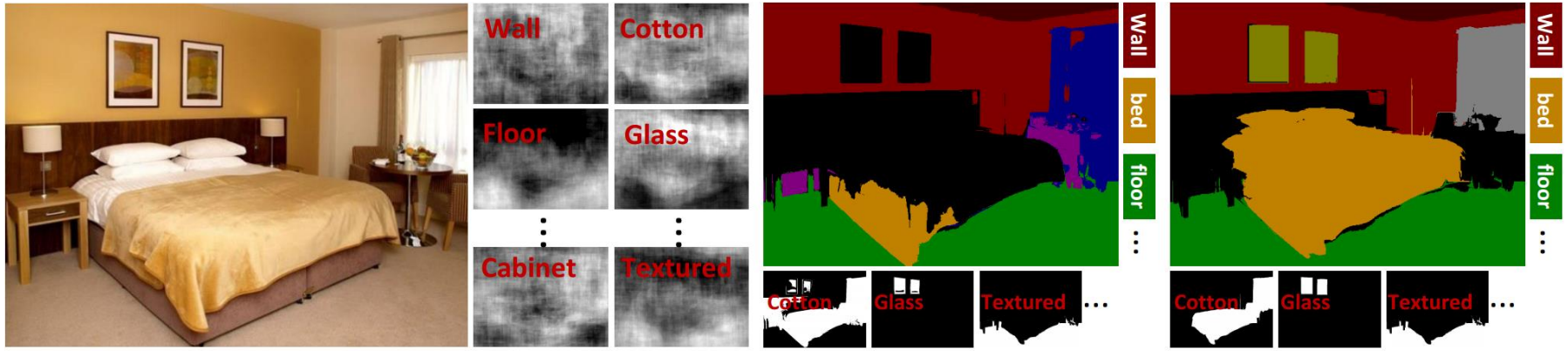
Make the wood cabinet in bottom-middle lower

Verbal guided image parsing

Make the wood cabinet in bottom-middle lower

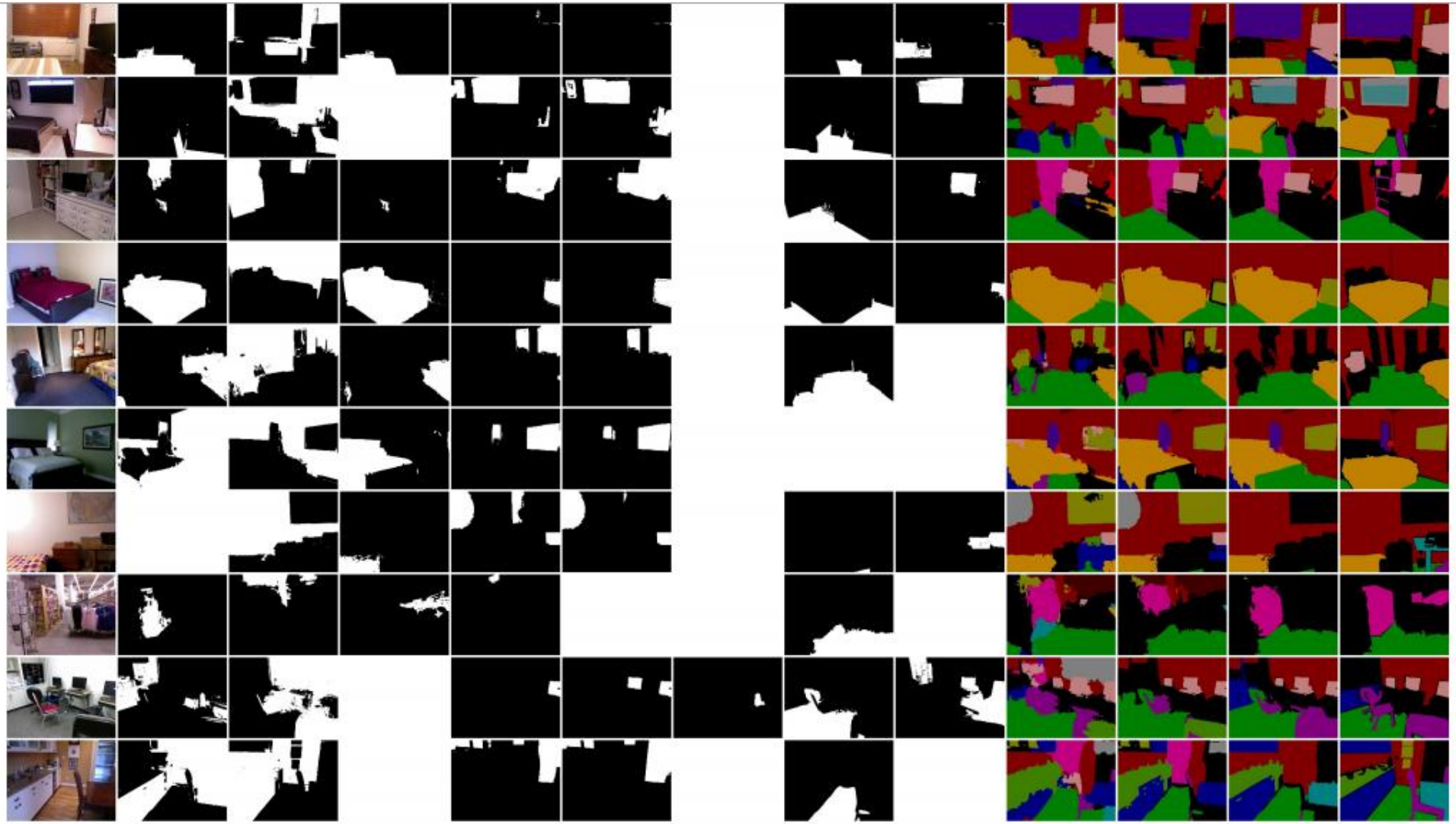


Verbal Guided Image Parsing



(a) Inputs: an image and learned weak hypothesis [Shotton et al. 2009] (b) Automatic scene parsing results (c) Natural language guided parsing

Verbal Guided Image Parsing Results



Source wood painted cotton glass glossy plastic textured shiny DenseCRF Our-auto V-Guided G-Truth

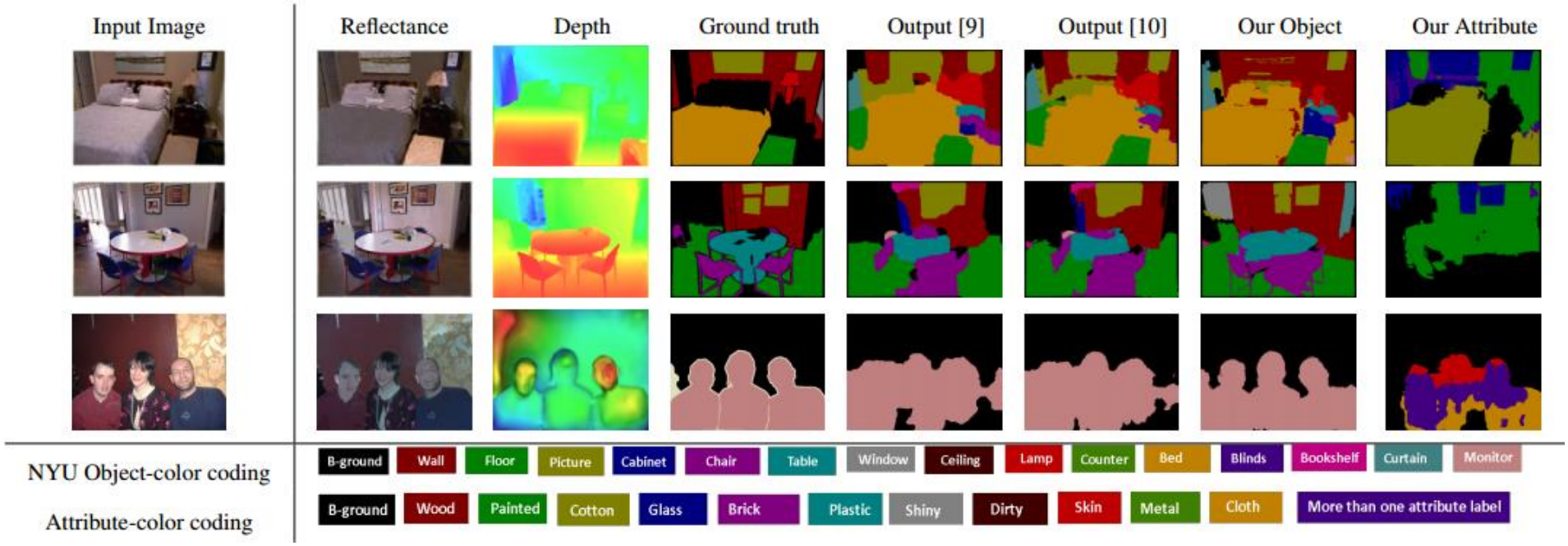
Discussion

- Pros
  - Empower users to interact with images through verbal commands.
  - Provide a potential crowd-sourcing way to collect image parsing annotations.
- Cons and future works
  - Limited supported commands and grammars
    - This would be further improved by exploring new tools, e.g. Recurrent Neural Networks (RNN).
  - Limited visual attributes
    - This would be advanced, by considering intrinsic images, e.g. OpenSurface Project, and Intrinsic Image in the wild.

Thank you!

http://www.robots.ox.ac.uk/~szheng/

Higher Order Priors for Joint Intrinsic Image, Objects, and Attributes Estimation, NIPS 2013.



Demo

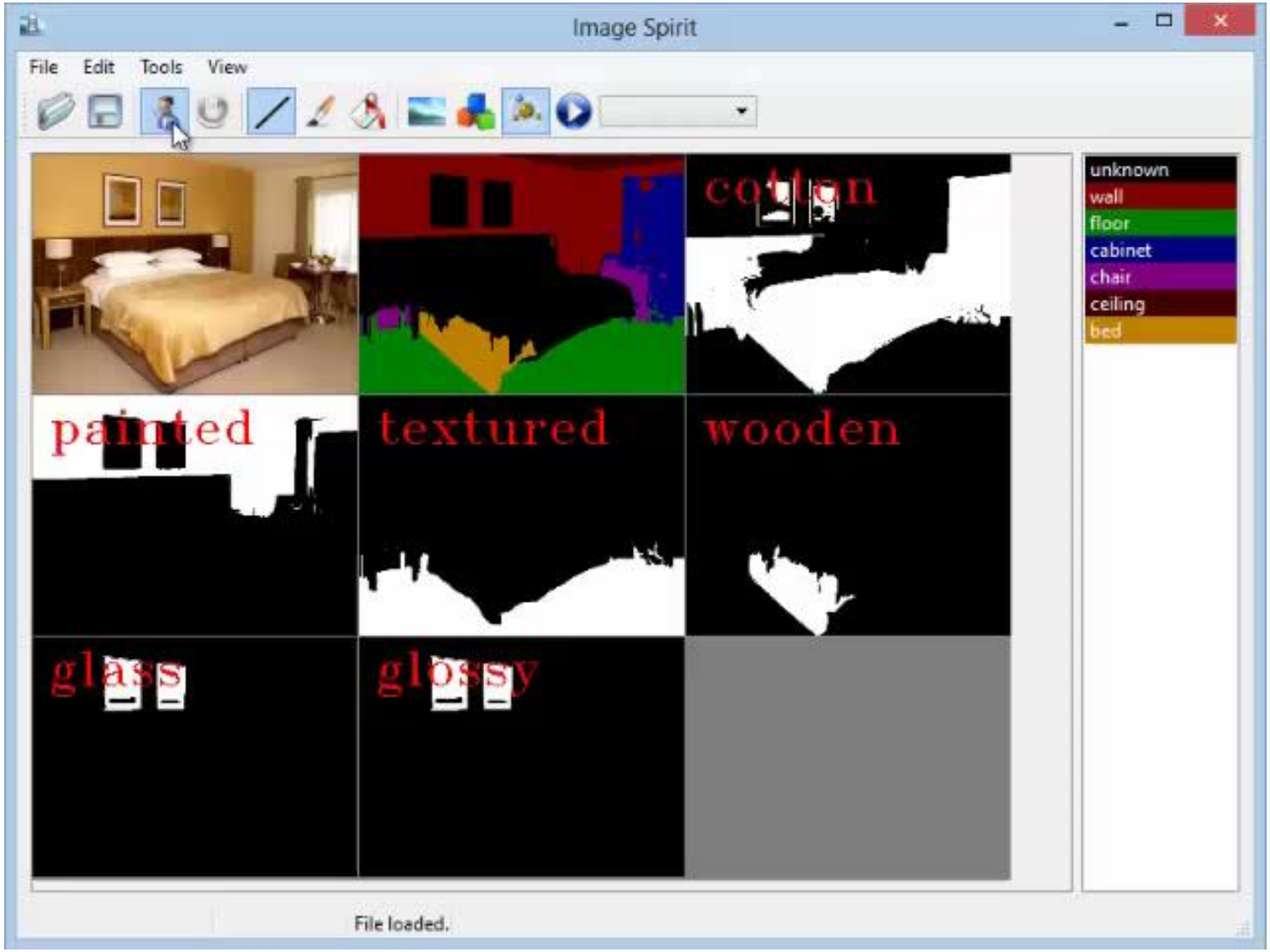


Image Spirit
File
Edit
Tools
View
cotton
painted
textured
wooden
glass
glossy
unknown
wall
floor
cabinet
chair
ceiling
bed
File loaded.