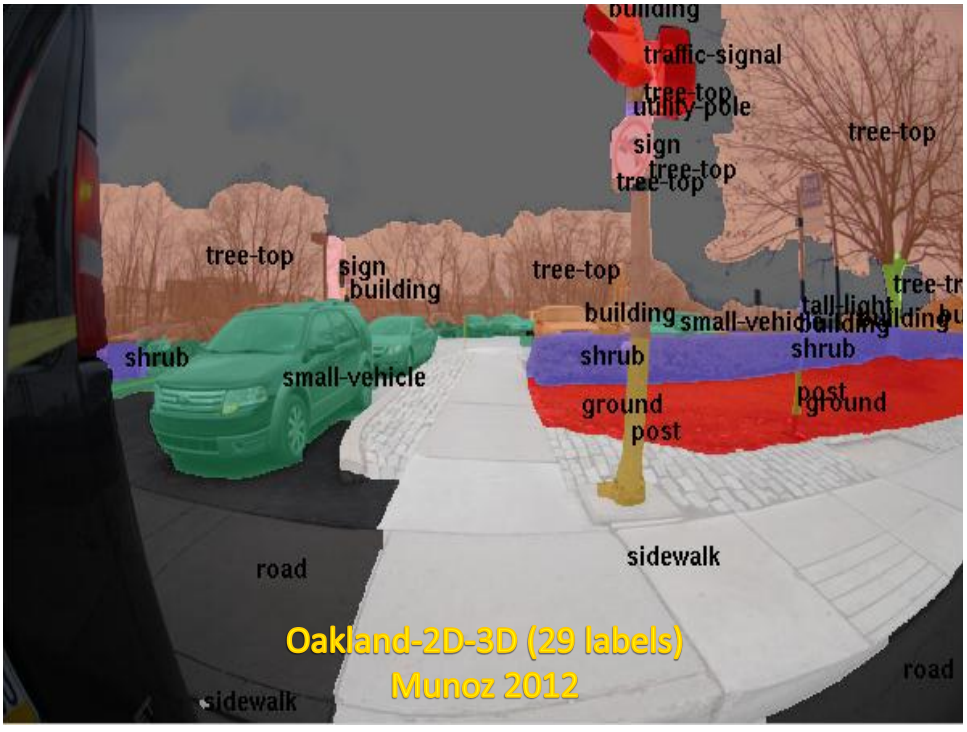
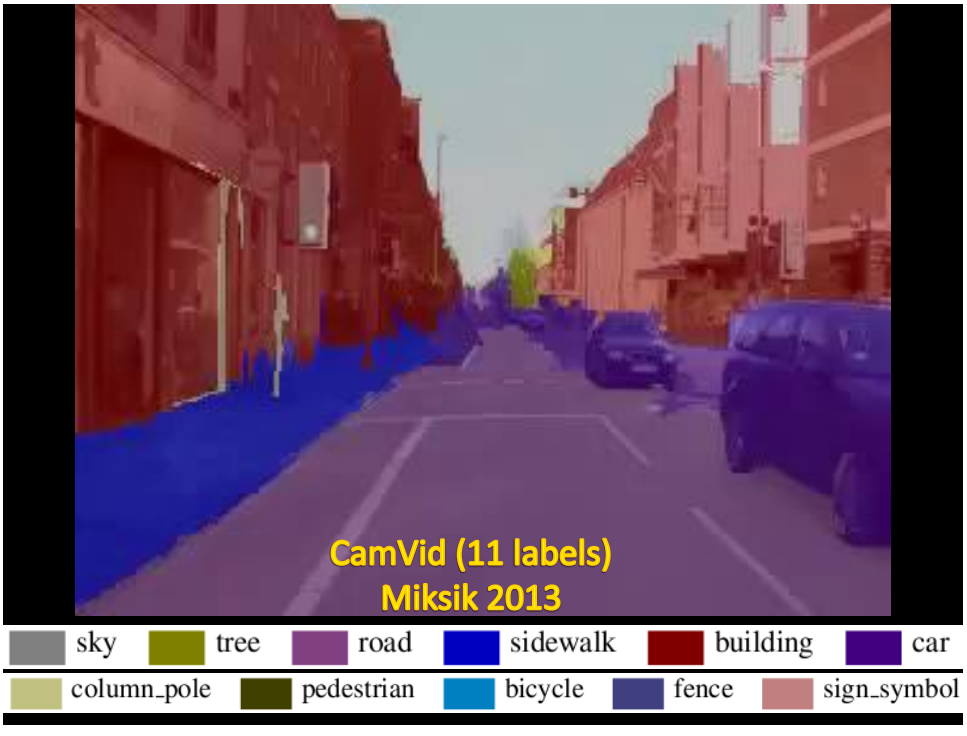
Semantic labeling

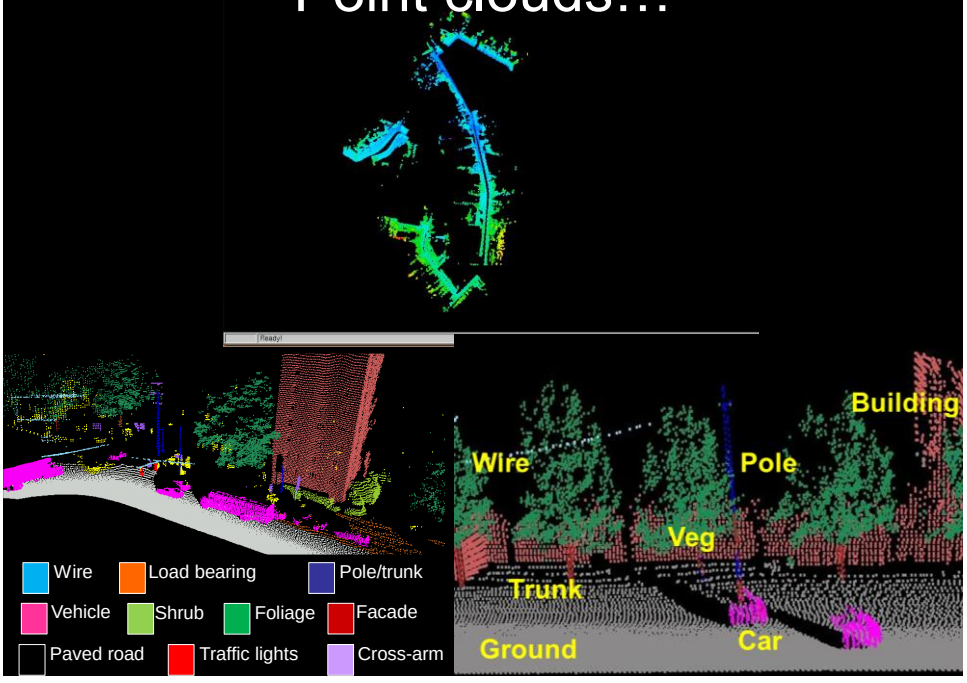
Martial Hebert, Drew Bagnell, Daniel Munoz
Carnegie Mellon University

www.cs.cmu.edu/~dmunoz/projects/
www.youtube.com/user/dmunozcmuri


building
traffic-signal
tree-top
utility-pole
tree-top
sign
tree-top
tree-top
tree-top
tree-top
sign
building
building
small-vehicle
shrub
shrub
shrub
small-vehicle
ground
post
ground
post
road
sidewalk
road
sidewalk
Oakland-2D-3D (29 labels)
Munoz 2012


CamVid (11 labels)
Miksik 2013
sky
tree
road
sidewalk
building
car
column_pole
pedestrian
bicycle
fence
sign_symbol

Point clouds…



1. Independent predictions

$$y^* = \arg\max_y \sum_i \phi(y_i, x)$$

1. Independent predictions

$$y^* = \arg\max_y \sum_i \phi(y_i, x)$$

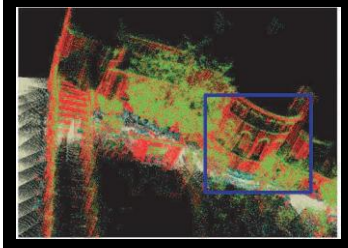
$$P(y|x) = \frac{1}{Z} \prod_i e^{\phi(y_i, x)}$$

1. Independent predictions

$$y^* = \arg\max_y \sum_i \phi(y_i, x)$$

- Features: local shape, texture, color, etc.
- Predictor: SVM, MaxEnt, etc.

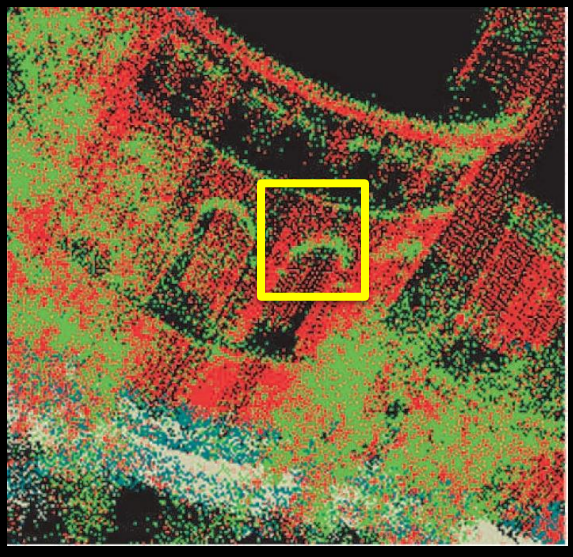
Needs context



Buildings

Tree Veg

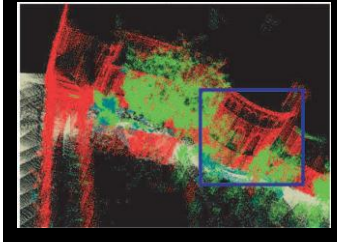
Shrubs



Anguelov 2005

No context

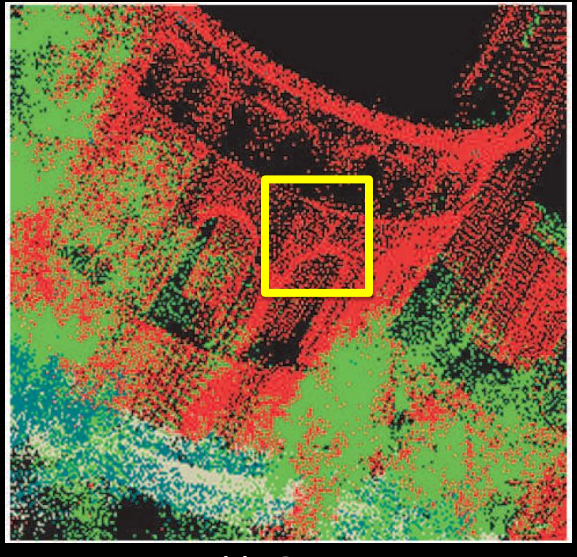
Needs context



Buildings

Tree Veg

Shrubs



Anguelov 2005

With Context

2. Pairwise context

$$y^* = \arg\max_y \sum_i \phi(y_i, x) + \sum_{i,j} \phi(y_i, y_j, x)$$

2. Pairwise context

$$y^* = \arg\max_y \sum_i \phi(y_i, x) + \sum_{i,j} \phi(y_i, y_j, x)$$

$$P(y|x) = \frac{1}{Z} \prod_i e^{\phi(y_i, x)} \prod_{i,j} e^{\phi(y_i, y_j, x)}$$

2. Pairwise context

$$y^* = \arg\max_y \sum_i \phi(y_i, x) + \sum_{i,j} \phi(y_i, y_j, x)$$

- Inference with BP, MCMC, LP, Graph-cuts
- NP-hard except in very special cases → Approximate inference only

Difficult from local context

Easier from broader context

Ideal regions

Fig. from Tomasz Malisiewicz

The reality

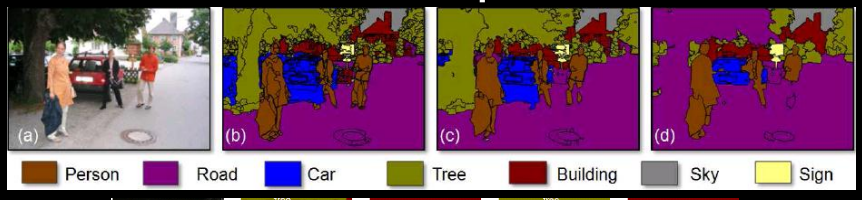
- How to represent potentials over larger support?
- How to deal with uncertainty on the choice of support regions? Multiple segmentations, hierarchical representations.

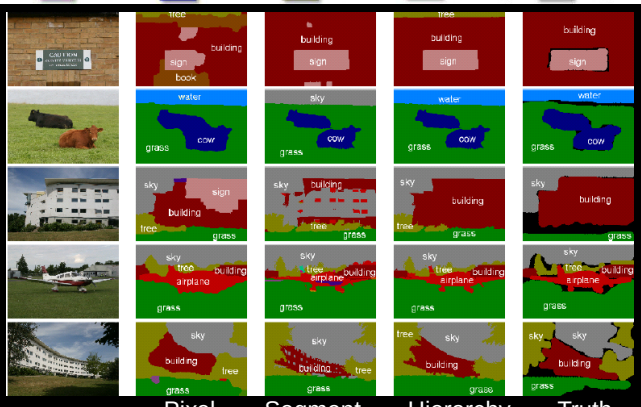
3. High-order context

$$y^* = \arg\max_y \sum_i \phi(y_i, x) + \sum_{i,j} \phi(y_i, y_j, x) + \sum_C \phi(y_c, x)$$

Graph cuts [Kohli 2010], MM [Taskar 2004, Munoz 2009], sampling [Gould 2010], ….

Example

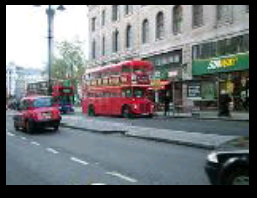




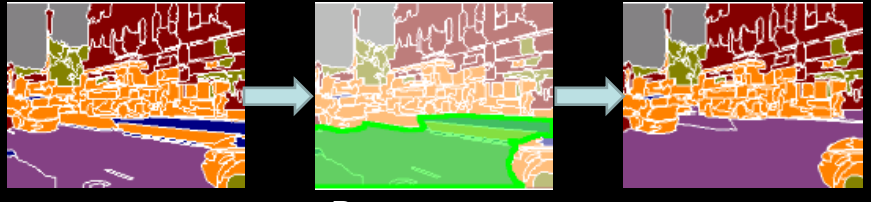
Pixel Segment Hierarchy Truth

L. Ladický, C. Russell, P. Kohli, P. H.S. Torr. Associative Hierarchical CRFs for Object Class Image Segmentation. International Conference on Computer Vision, 2009.

Example: Sampling



Multiple segmentations



Propose move

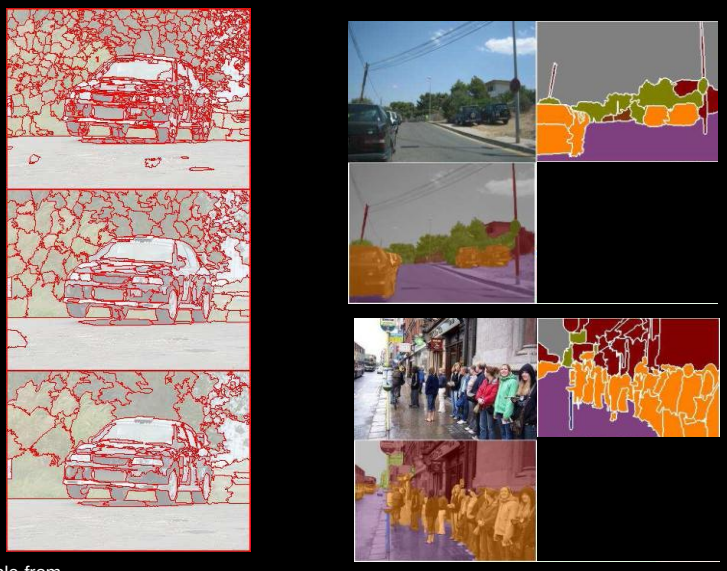
$E(y|x)$

Evaluate energy

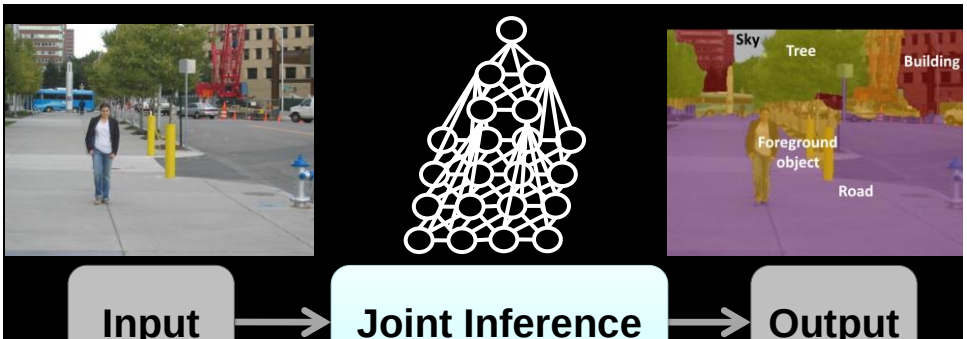
Accept if move reduces energy

[Example from Gould et al.]

Gould, S., Fulton, R., Koller, D.: Decomposing a scene into geometric and semantically consistent regions. In: ICCV (2009)
Stephen Gould, Tianshi Gao and Daphne Koller. Region-based Segmentation and Object Detection. In Advances in Neural Information Processing Systems (NIPS), 20010



[Example from Gould et al.]

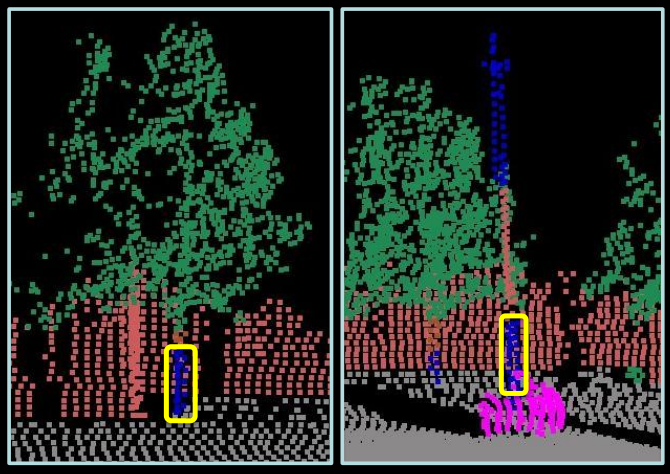


$$y^* = \underset{y}{\operatorname{argmax}}\, f(x, y)$$

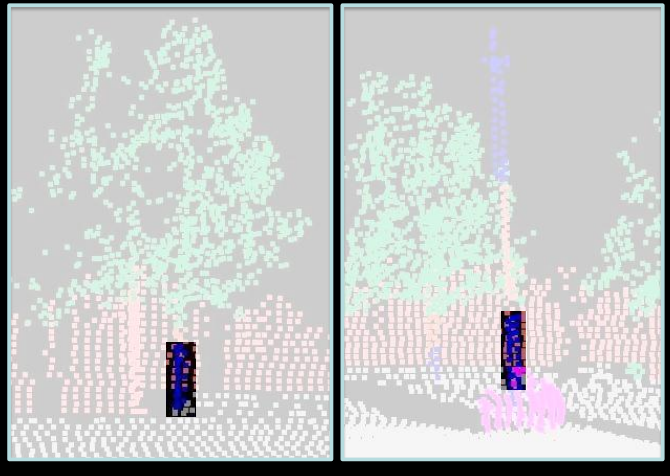
Training: Model the distribution of labels vs. features
Inference: Find most likely labels given model

[Bouman 1994, Feng 2002, He 2004, Borenstein 2004, Kumar 2005, Ladicky 2010, Gould 2010., Shotton et al., Zhu 2006, Yao 2010, Ren 2012 …]

1. Limited interactions

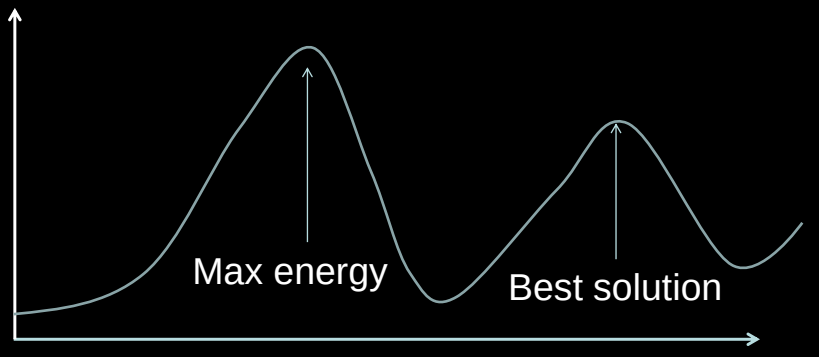


1. Limited interactions



- Restrictive relations among regions [Ladicky 2009]
- Restrictive potentials over regions [Kohli 2010]

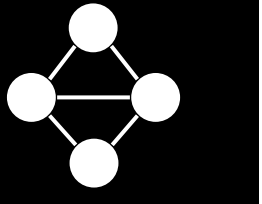
2. Model mismatch



- Analysis in the context of stereo [Szeliski 2008]
- Matched learning and inference [Kumar 2005, Wainwright 2006, Ranzato 2007, Lowd 2008]

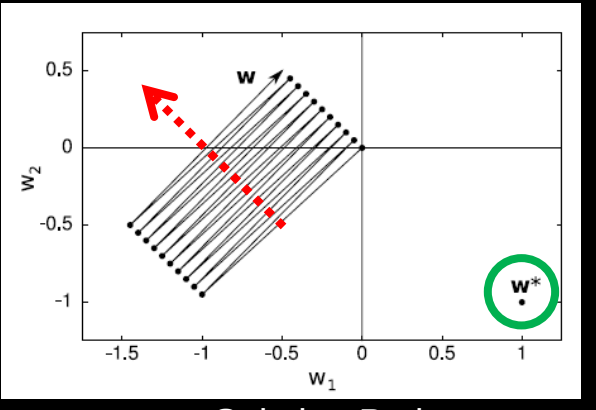
3. Provably hard to learn

Wainwright 2006, Kulesza and Pereira 2008, Finley and Joachims 2008



Simple Graphical Model

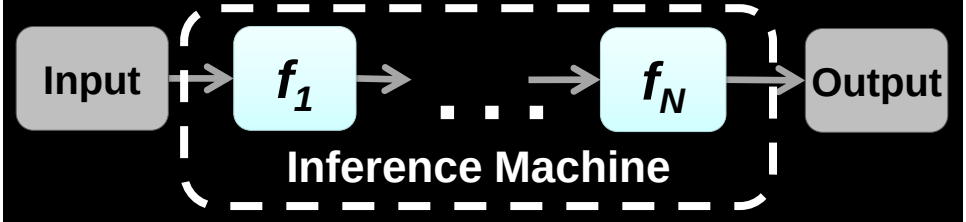
$w = \begin{bmatrix} w_1 \\ w_2 \end{bmatrix}$



Solution Path

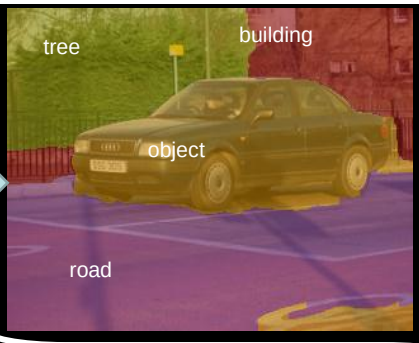
Example from Kulesza and Pereira 2008

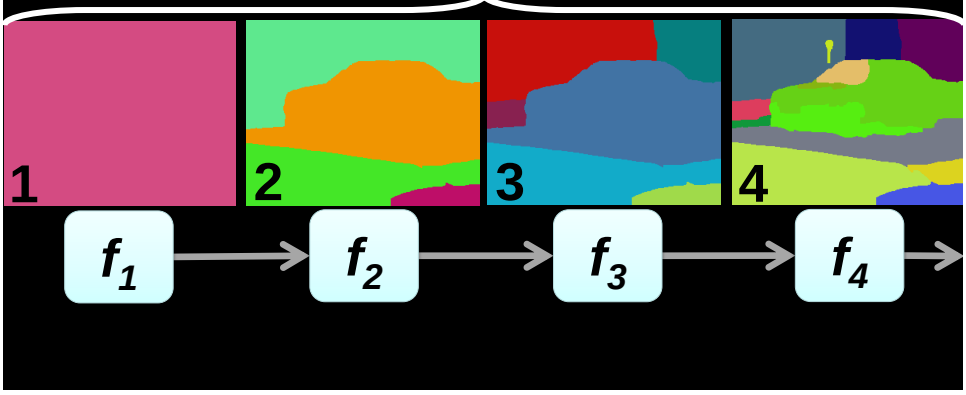
HIM: Hierarchical Inference Machine

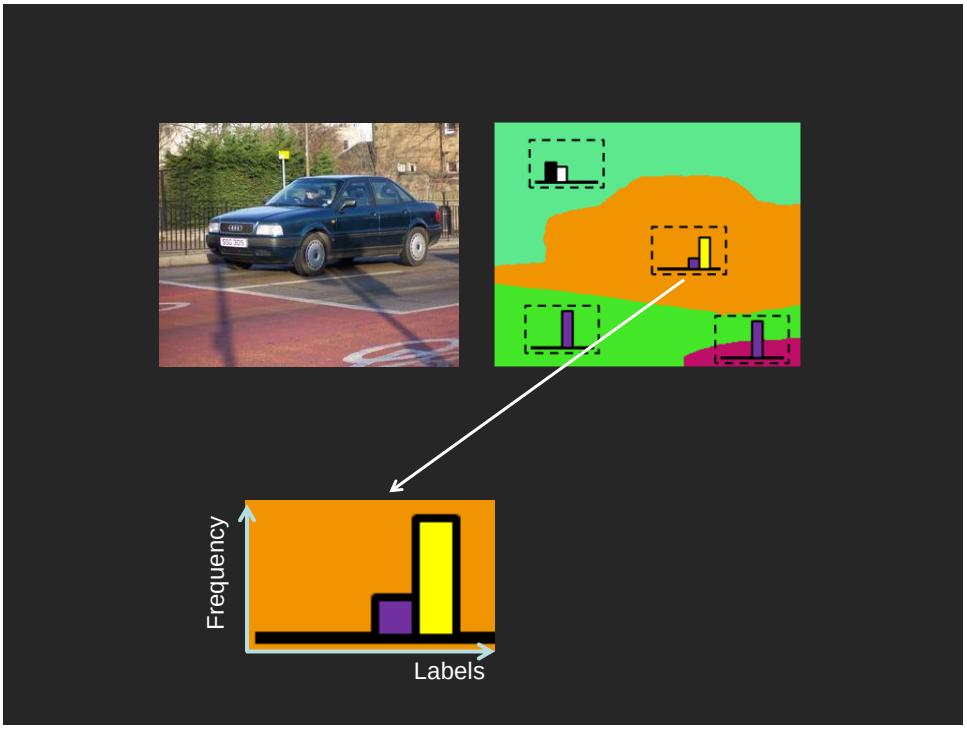


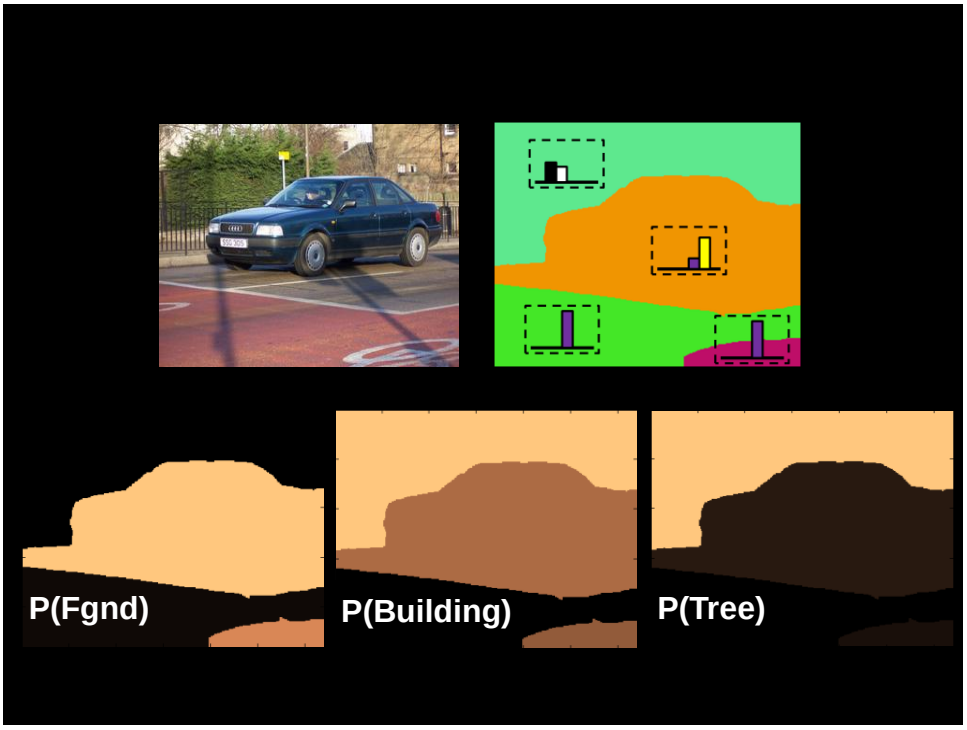
Do not learn a joint distribution/energy

Instead, learn a *procedure* designed to *iteratively* decode the scene






Frequency
Labels


P(Fgnd)
P(Building)
P(Tree)

Level 1/8 Predictions

Segmentation

Level 1/8 Predictions

Segmentation

P(Foreground)

15%

P(Tree)

P(Building)

P(Road)

18%

12%

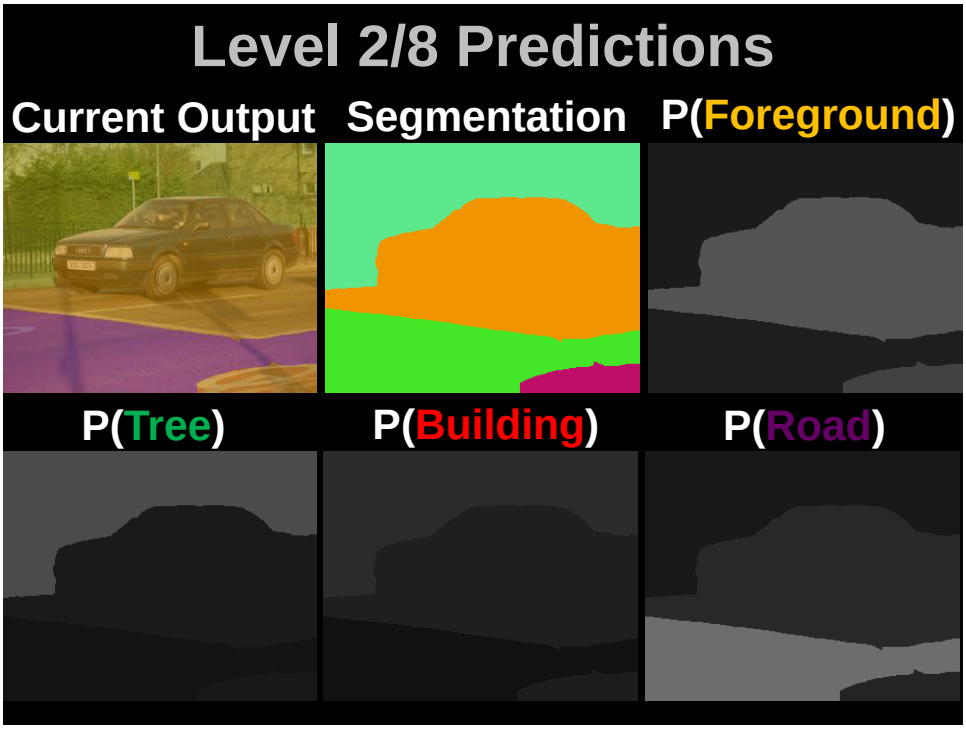
31%

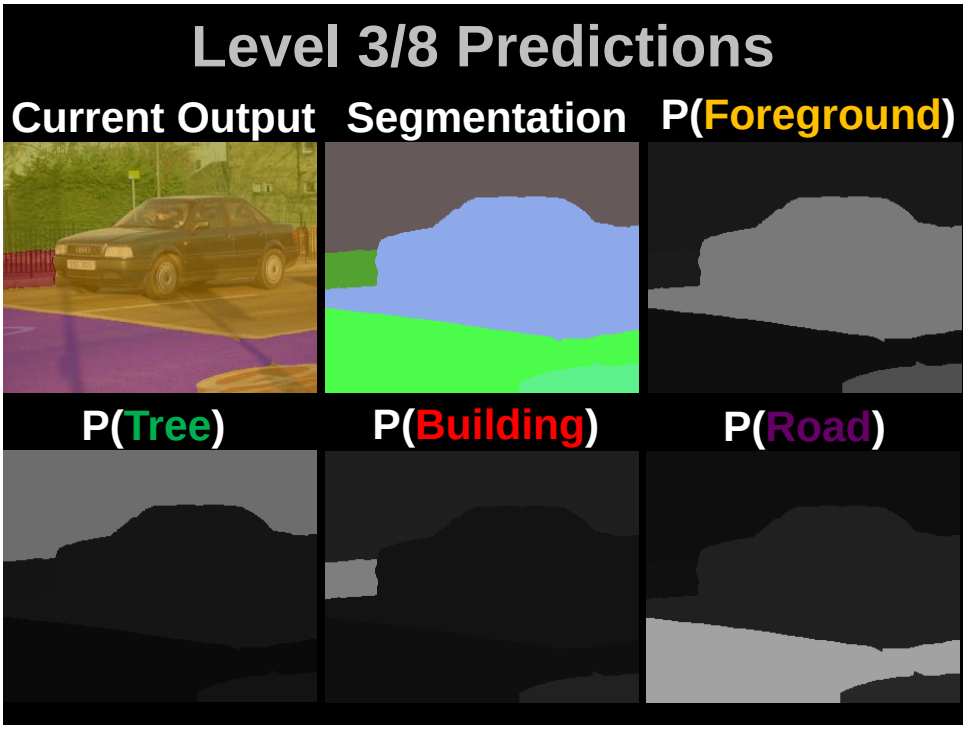
Level 1/8 Predictions

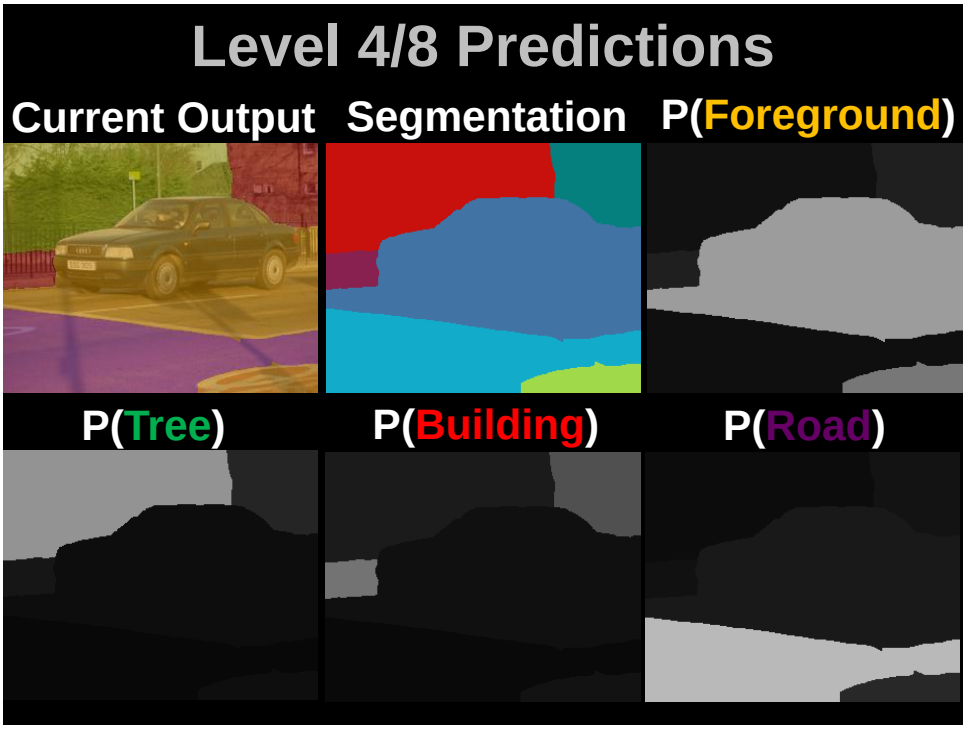
Current Output	Segmentation	P(Foreground)
		15%
P(Tree)	**P(Building)**	**P(Road)**
18%	12%	31%

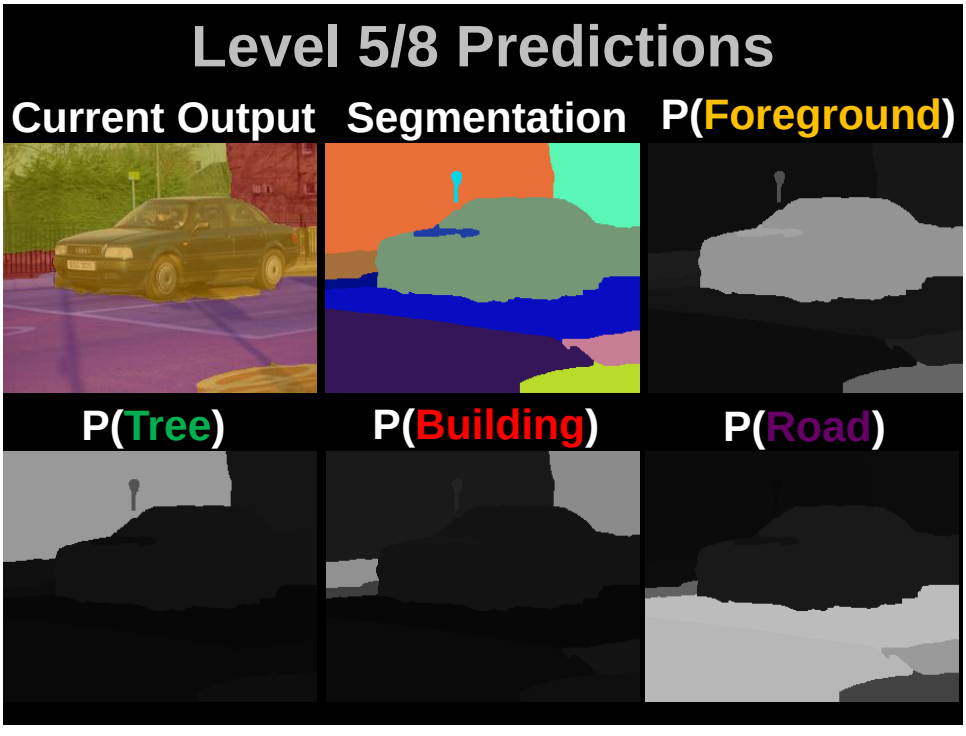
Level 2/8 Predictions

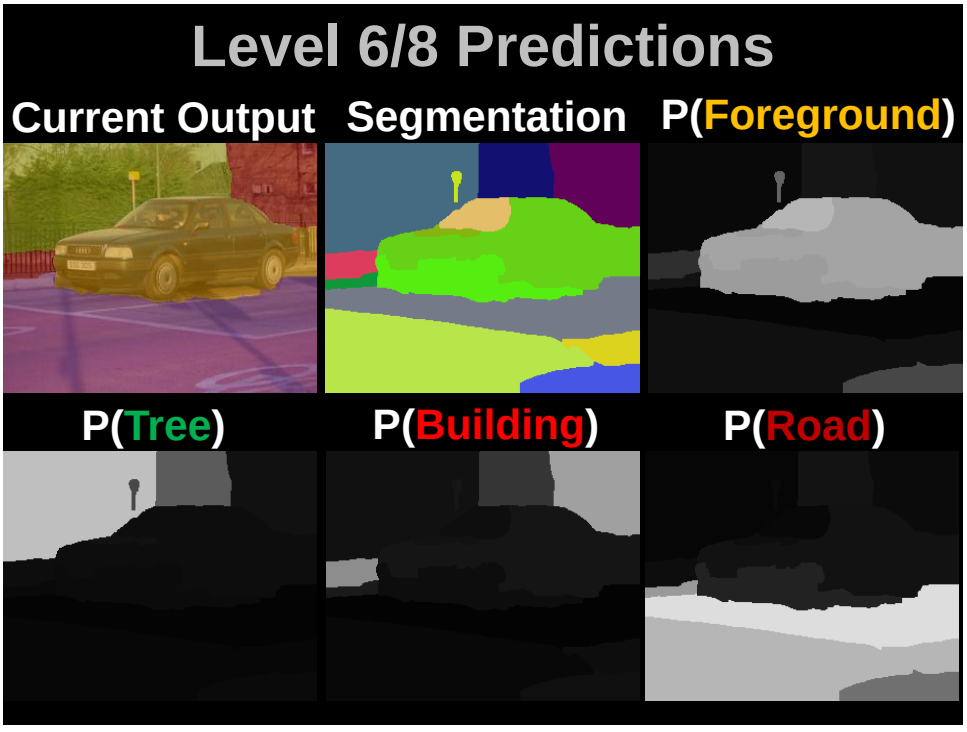
	Segmentation	P(Foreground)
		
P(Tree)	**P(Building)**	**P(Road)**
		

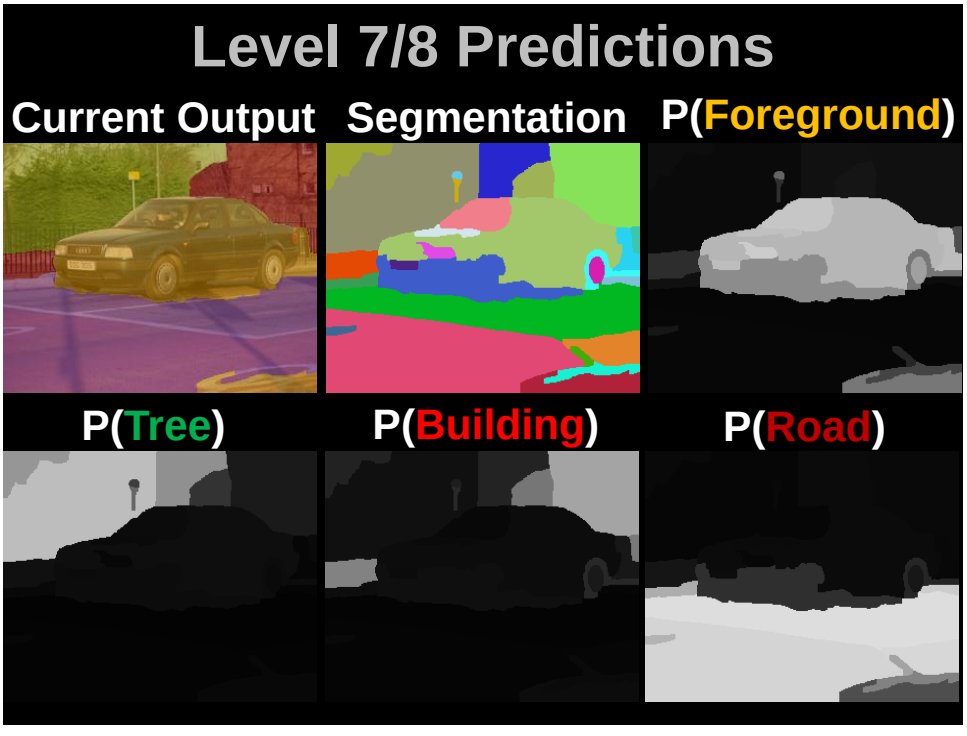

Level 2/8 Predictions
Current Output
Segmentation
P(Foreground)
P(Tree)
P(Building)
P(Road)

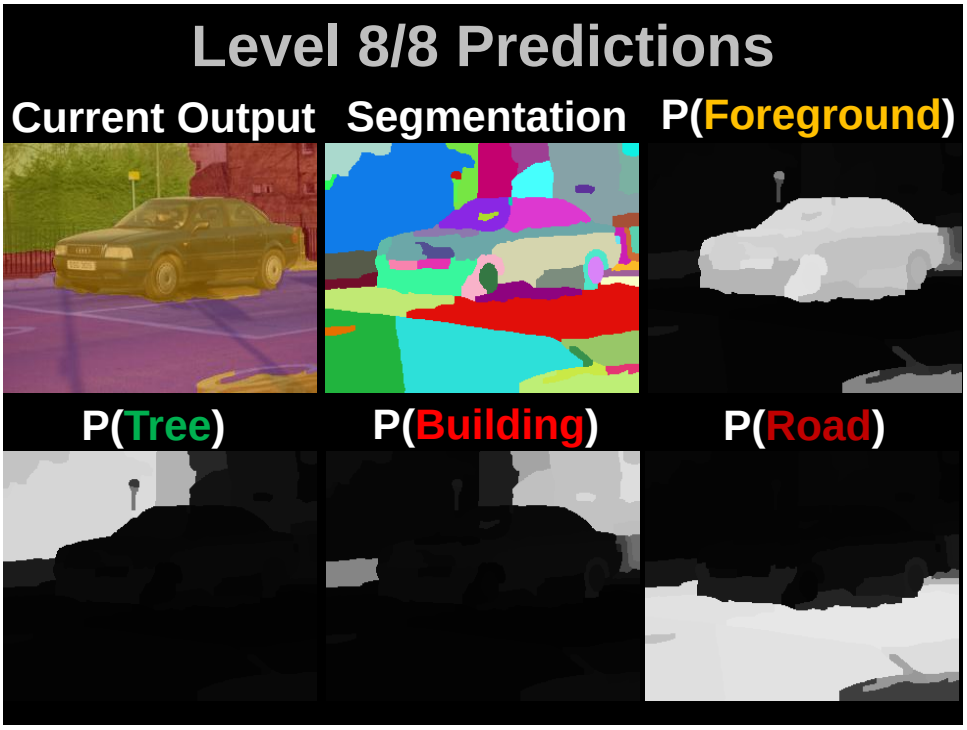

Level 3/8 Predictions
Current Output
Segmentation
P(Foreground)
P(Tree)
P(Building)
P(Road)

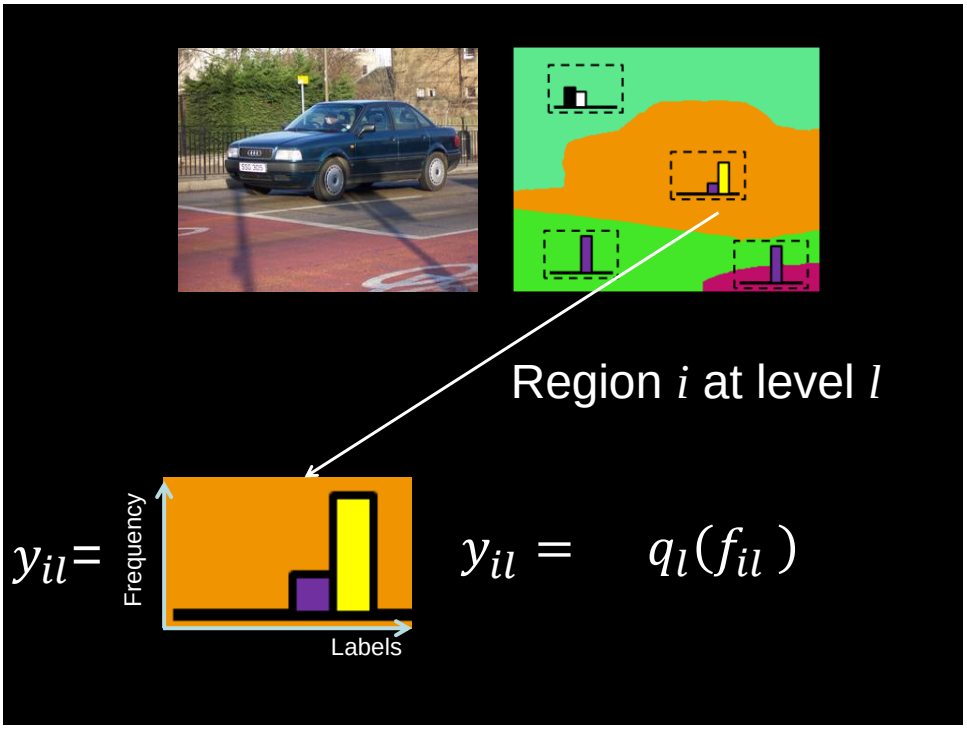

Level 4/8 Predictions
Current Output
Segmentation
P(Foreground)
P(Tree)
P(Building)
P(Road)

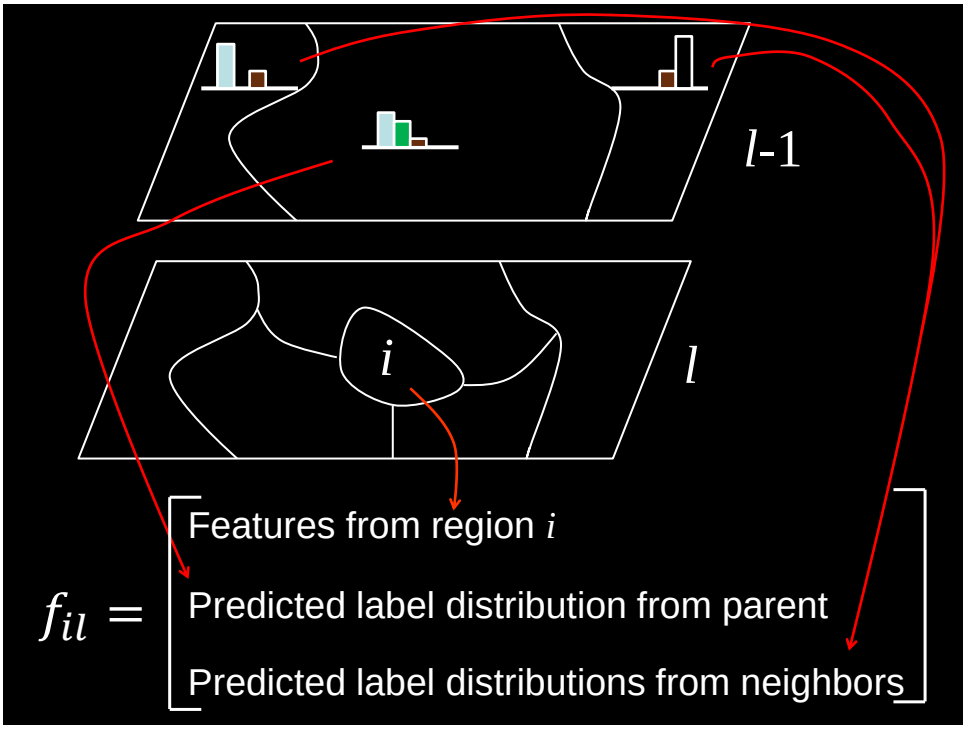

Level 5/8 Predictions
Current Output
Segmentation
P(Foreground)
P(Tree)
P(Building)
P(Road)

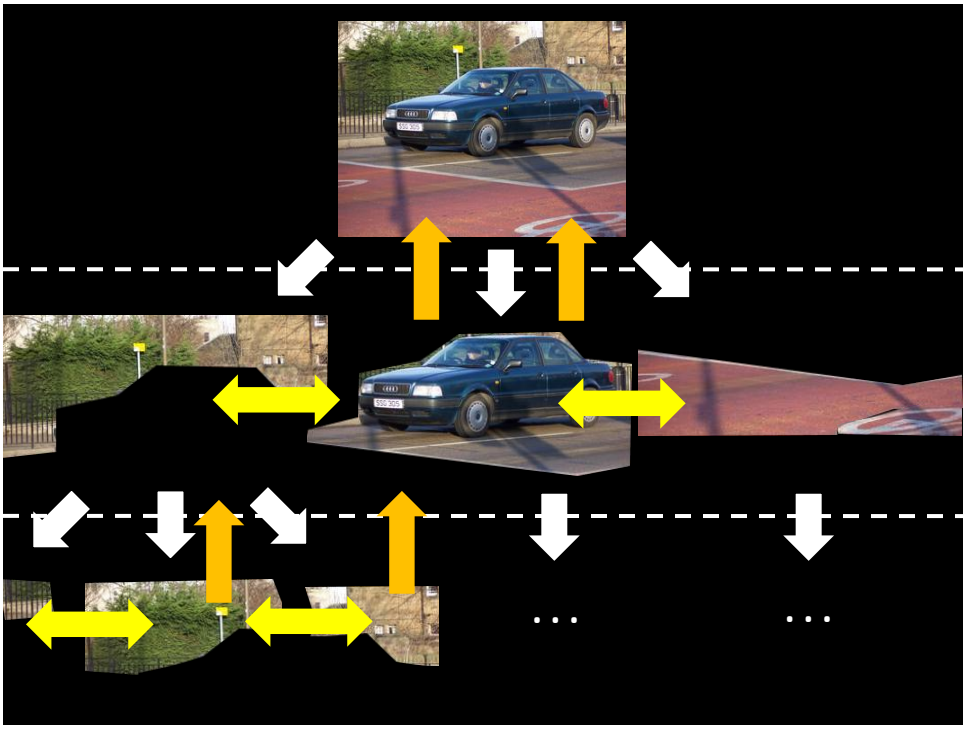

Level 6/8 Predictions
Current Output
Segmentation
P(Foreground)
P(Tree)
P(Building)
P(Road)

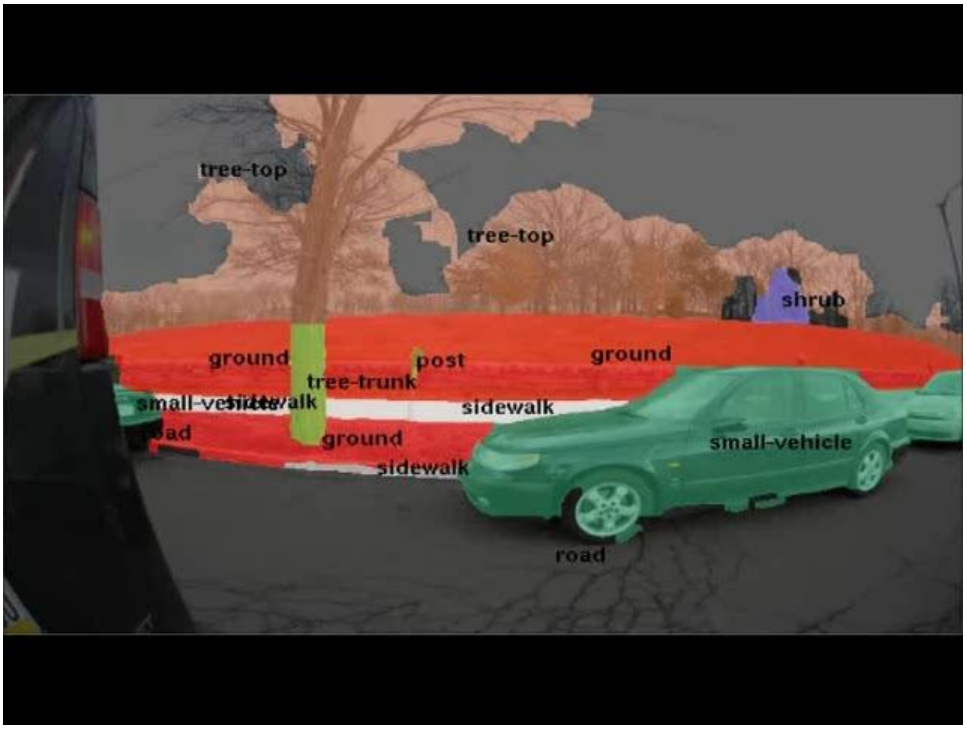

Level 7/8 Predictions
Current Output
Segmentation
P(Foreground)
P(Tree)
P(Building)
P(Road)


Level 8/8 Predictions
Current Output
Segmentation
P(Foreground)
P(Tree)
P(Building)
P(Road)


Region i at level l
y_{il}=
Frequency
Labels
$y_{il} = q_l(f_{il})$


l-1
i
l
$f_{il} =$
Features from region i
Predicted label distribution from parent
Predicted label distributions from neighbors

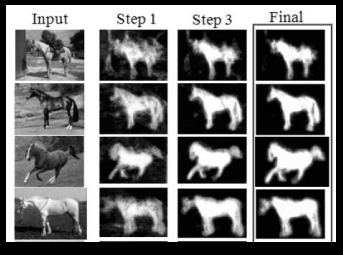

…
…



Connections

- Auto-context

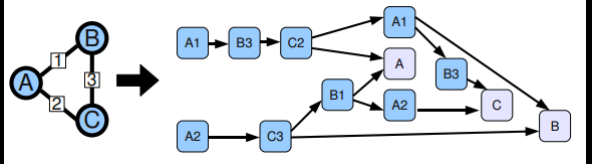
Z. Tu and X. Bai. Auto-context and Its Application to High-level Vision Tasks. PAMI, 32(10), 2010.



- Inference machines

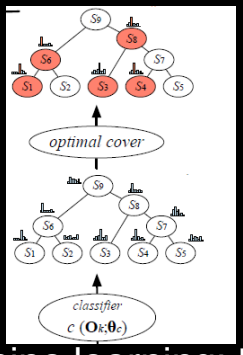
S. Ross, D. Munoz, M. Hebert, and J.A. Bagnell. Learning Message-Passing Inference Machines for Structured Prediction CVPR 2011.

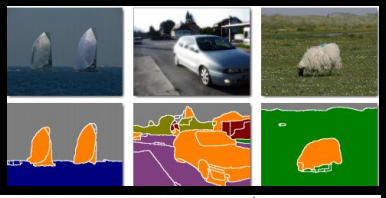
R. Shapovalov, D. Vetrov, P. Kohli. Spatial Inference Machines. CVPR 2013.

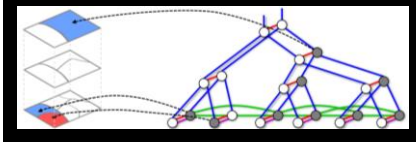


Connections

- Hierarchies
  - C. Farabet, C. Couprie, L. Najman, Y. LeCun. Learning hierarchical features for scene labeling. PAMI 2012.
  - V. Lempitsky, A. Vedaldi, and A. Zisserman. A Pylon Model for Semantic Segmentation. NIPS 2011.

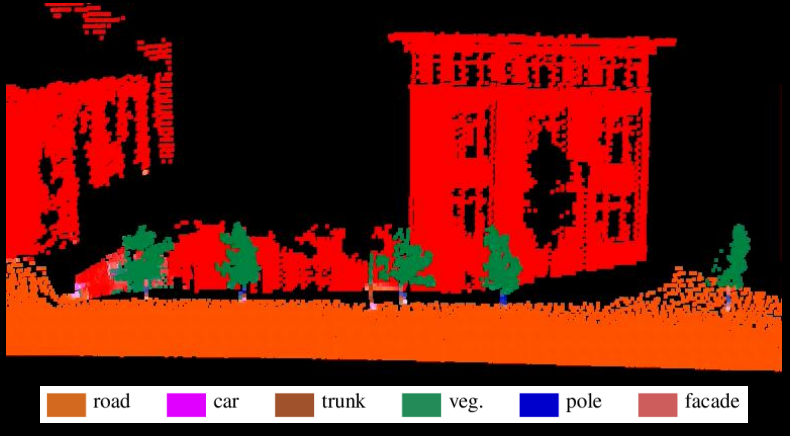






- Machine learning: Reductions and search
  - H. Daume III, J. Langford, and D. Marcu. Search-based Structured Prediction. Machine Learning, 75(3), 2009.
  - S. Ross and J. A. Bagnell. Efficient Reductions for Imitation Learning. IASTAT, 2010.
  - Langford et al. Machine Learning Reductions Tutorial. ICML 2009.

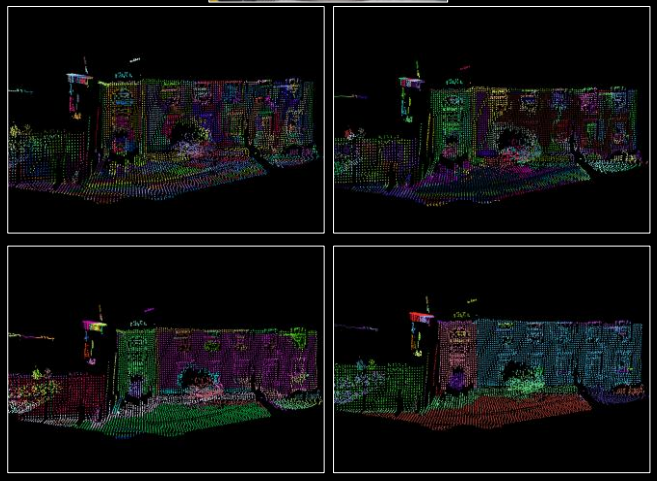
Extensions



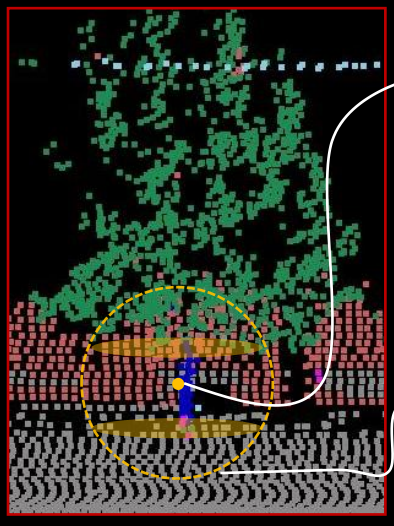
Point clouds….

Exactly the same approach with 3D segmentation and 3D features

Segmentation

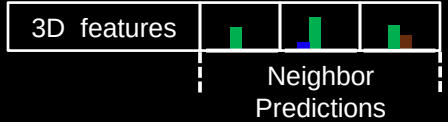


Features

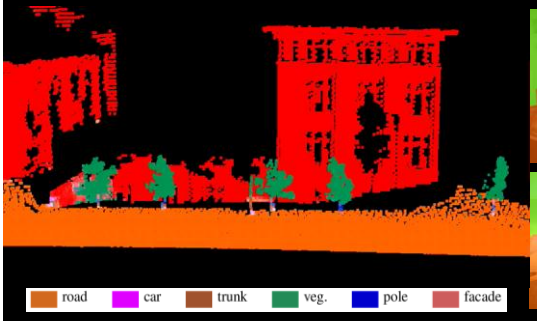


- Quantized with soft k-means [Coates AISTASTS 2011]:
  - Spectral features at 2 scales
  - Spin images
- Relative elevation map
- XYZ stats
- 97-dim

- Distribution of labels in neighborhood
- Three layers (top, middle, bottom)

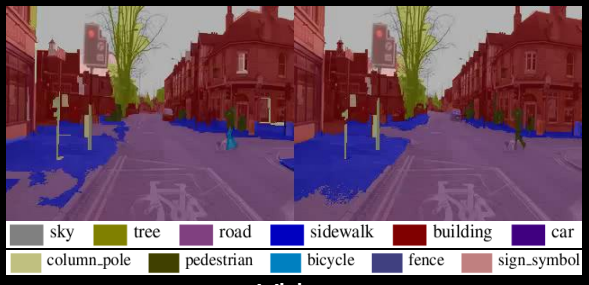


Extensions



Point clouds….

Other labeling tasks



Video

Learning issues

- Option 1: Train each level independently using ground truth labels from parent

f_1 f_2 f_3 f_4

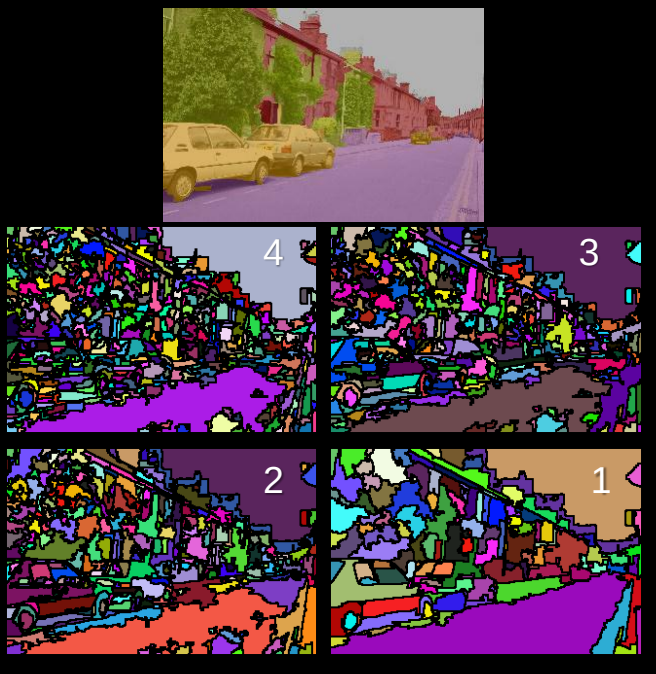
- Option 2: Train each level by using the predictions from the parent level using a single training set

$f_1 \rightarrow f_2 \rightarrow f_3 \rightarrow f_4 \rightarrow$

Details (Example)

- Features
  - Region geometry
  - Soft bags-of-words [Coates AISTATS 2011] over per-pixel features
    - Texture, LBP, SIFT/HoG [Ladicky ECCV 2010]
  - 914-dim descriptor
- Segmentation:
  - Felzenszwalb-Huttenlocher graph segmentation modified at multiple levels by varying parameters (τ and min regions)

LBP: 8 comp
Col SIFT
16: G,GD,L



Details (Example)

- Features
  - Region geometry
  - Soft bags-of-words [Coates AISTATS 2011] over per-pixel features
    - Texture, LBP, SIFT [Ladicky ECCV 2010]
- Segmentation:
  - Graph segmentation modified at four levels by varying the parameters (τ and min regions)
- Prediction:
  - Boosted vector-output regression trees
  - Stacked training [Cohen IJCAI 2005]
  - Prediction order:
    - root → leaves → root
    - 2 iterations at each level

LBP: 8 comp
Col SIFT
16: G,GD,L

Stanford Background Dataset

Gould ICCV 2009	76.4	65.5
Munoz ECCV 2010 (v1)	**76.9**	**66.2**
Tighe ECCV 2010	77.5	--
Socher ICML 2011	78.1	--
Lempitsky NIPS 2011	81.9	72.4
Ren CVPR 2012	82.9	74.5
Munoz ECCV 2010 (v2) **Segmentation**: 0.10,	**81.6** **Features**: 0.46,	**71.8** **Inference**: 0.04

Cambridge Video Dataset

Method	Overall Pixel	Average Class
Brostow ECCV 2008	69.1	53.0
Sturgess BMVC 2009	83.8	59.2
Zhang ECCV 2010	82.1	55.4
Ladicky ECCV 2010	83.8	62.5
de Nijs IROS 2012	75.0	54.7
Tighe IJCV 2013	83.3	51.2
Munoz ECCV 2010 (v2)	**85.7**	**60.5**

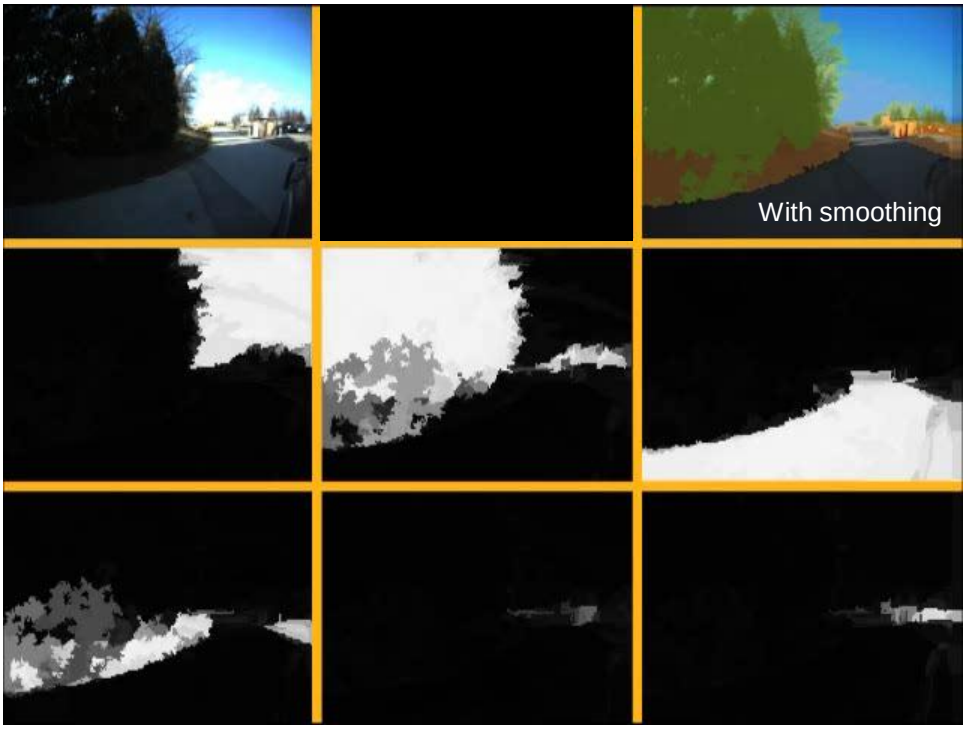
Digression: How to measure performance

Before After

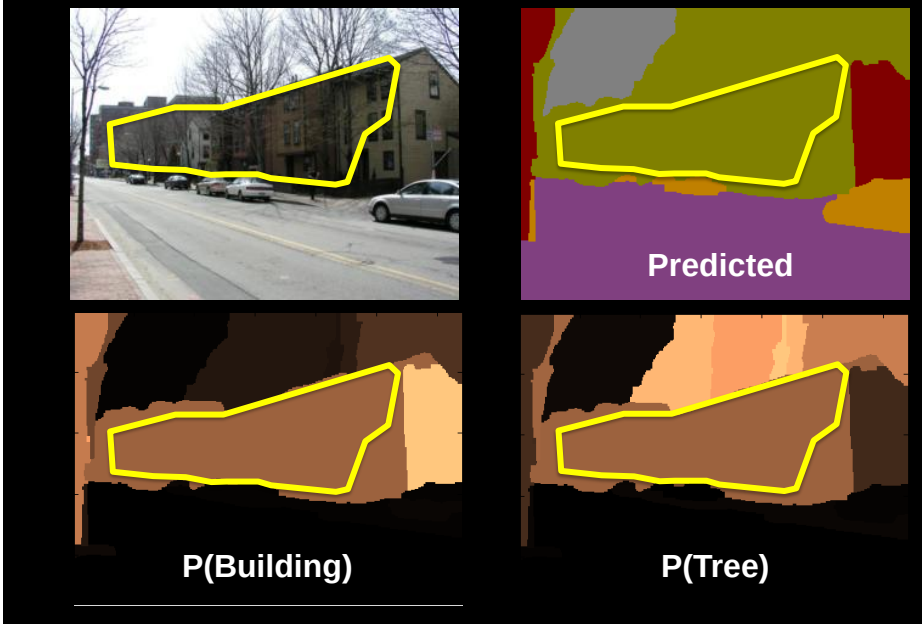
How to evaluate performance in a *task-relevant* way? Are accuracy, avg. precision, F1, useful?

Digression: The need for distributions

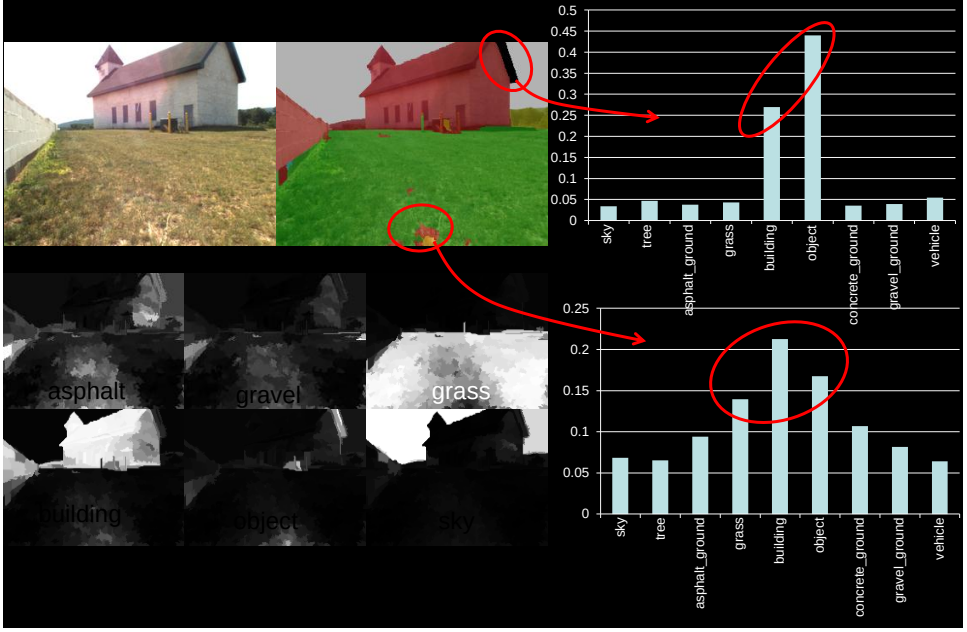
Visual input → Perception algorithms → Interpretation → Application

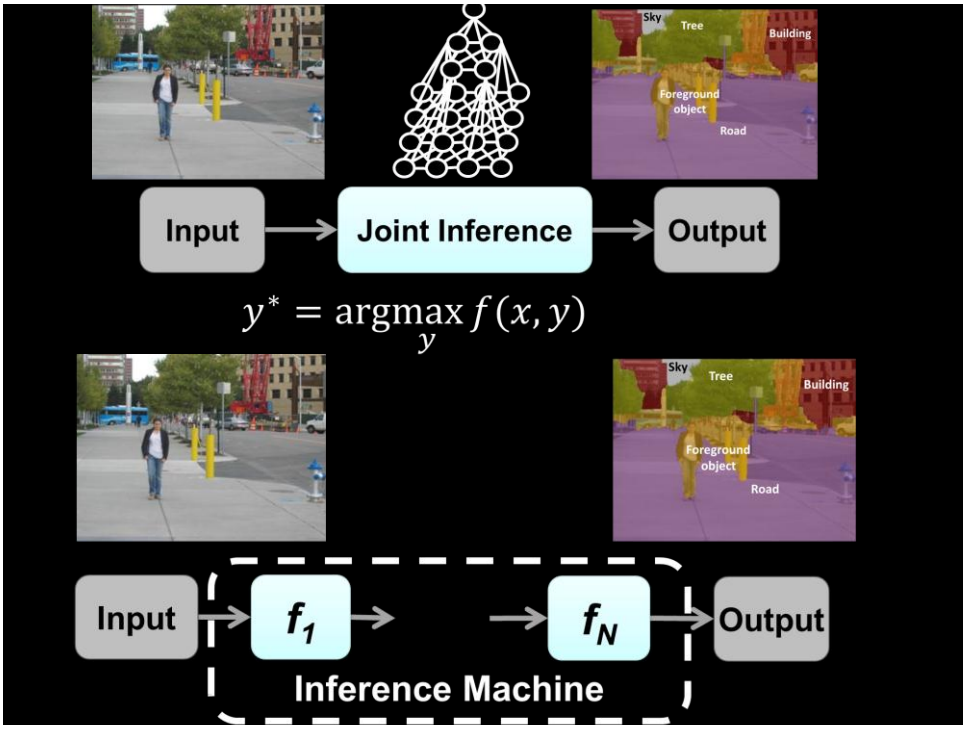


Representing ambiguity



Explaining decisions and mistakes




Sky
Tree
Building
Foreground object
Road
Input
Joint Inference
Output
$y^* = \underset{y}{\operatorname{argmax}} f(x, y)$
Sky
Tree
Building
Foreground object
Road
Input
f_1
f_N
Output
Inference Machine