# STOCHASTIC SEGMENTATION TREES

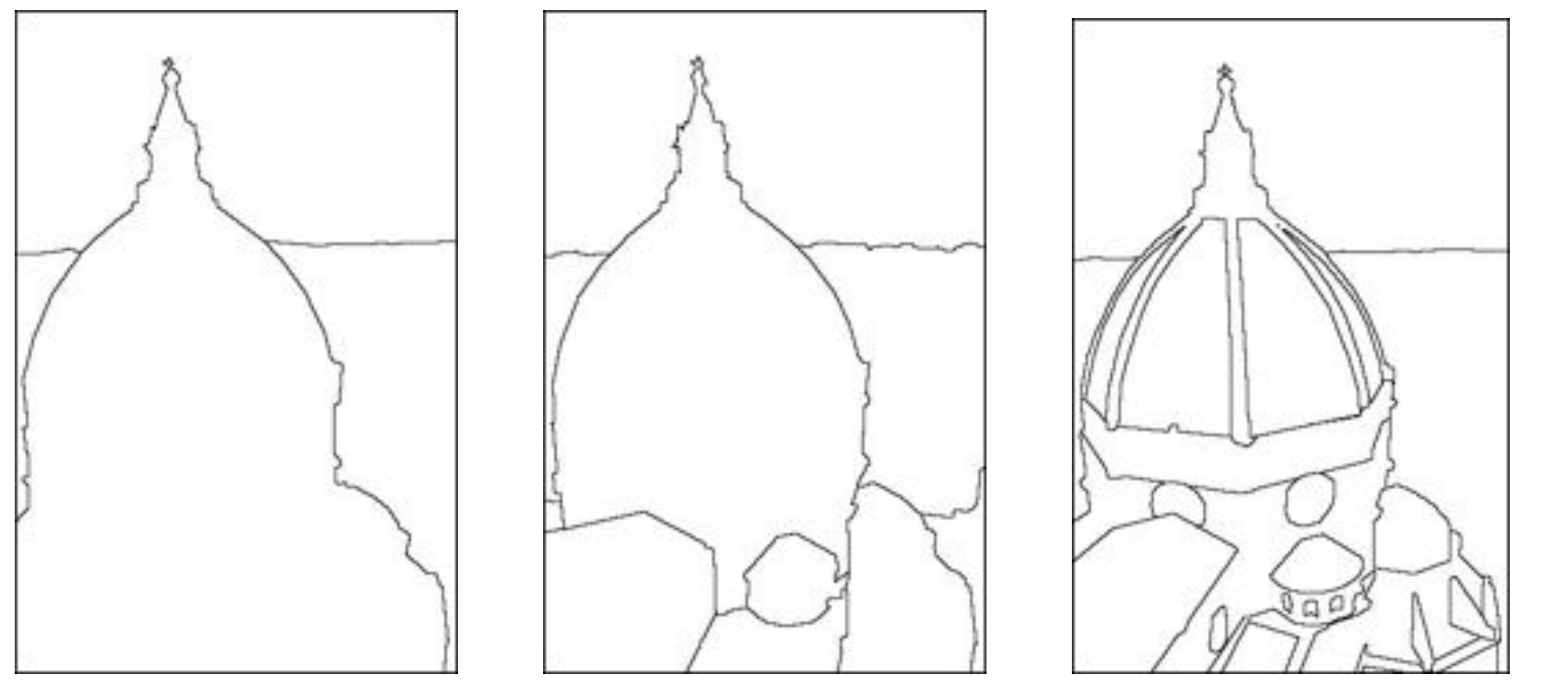
Snell Jake, Zemel Richard S.
University of Toronto
{jsnell,zemel}@cs.toronto.edu

## ABSTRACT

Many structured output problems such as image segmentation admit multiple correct outputs for a single input. We present a recursive neural network-based framework for modeling multiple output segmentations via a hierarchical tree of image regions. We perform learning by minimizing KL divergence from a target distribution constructed using a task-specific loss function from the ground truths. We conduct experiments on segmentations synthesized from the Penn-Fudan pedestrian dataset.

## GOAL AND MOTIVATION

An image from the Berkeley Segmentation dataset



Three of the corresponding ground truth segmentations

- Many structured output problems admit multiple correct outputs for a given input.
- Instead of reducing these to a single target, we want to capture the variations in outputs.
- **Goal**: Treat multiple ground truth segmentations as a target distribution while taking advantage of the hierarchical structure inherent to natural images. The model should predict multiple plausible outputs at test time.

## NOTATION

- $x$: input image
- $S = \{s_1, \ldots, s_M\}$: ground truth segmentations
- $\Delta(s_j, s) = 1 - RI(s_j, s)$: loss of predicting $s$ relative to ground truth $s_j$
- $RI(s_j, s)$ (Rand Index): sum of pixel pairs that have the same label in $s_j$ and $s$ and those that have different labels in both, divided by the number of pixel pairs
- $z$: a region hierarchy consisting of nodes $z_1, \ldots$
- $c_i$: feature representation of $z_i$
- $\theta$: model parameters
- $\mathcal{N}(z)$: non-terminal nodes of $z$
- $y_i$: binary label corresponding to node $z_i \in \mathcal{N}(z_i)$
- $\mathcal{Y}(z)$: set of binary labelings $y$ such that the label of a child is greater than or equal to that of its parent

## RNN-BASED REGION HIERARCHY

- Similar to RNN framework of (Socher et al. 2011)
- Tree formed by merging neighboring image region
- Start by extracting features from superpixels
- Each node $z_i$ has a fixed-length feature vector $c_i$
- Merges made by greedily maximizing a scoring function:

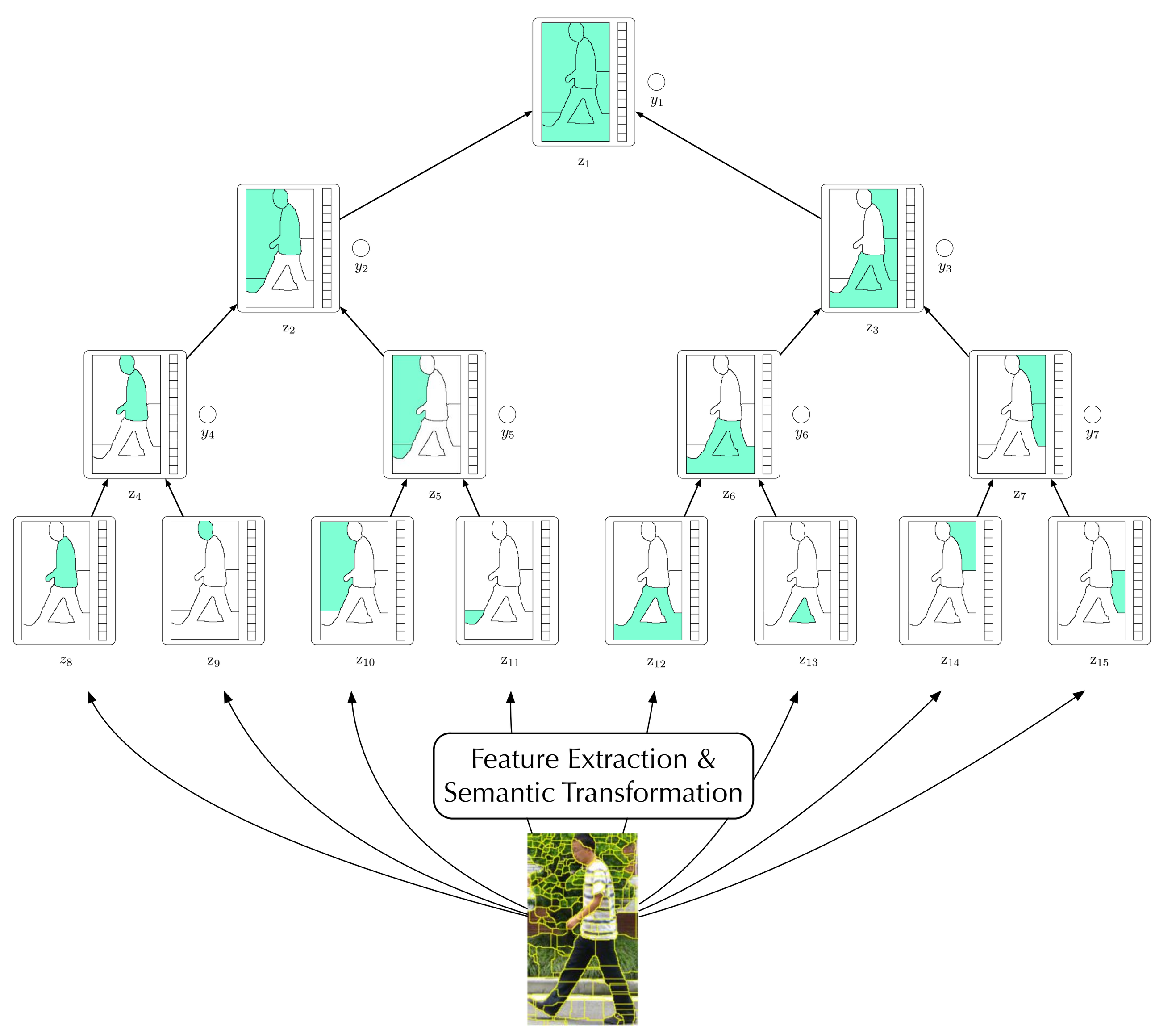
$$\psi_k = g^{\text{score}}(W^{\text{score}}[c_i; c_j] + b^{\text{score}})$$

- Parent feature vectors computed from children:

$$c_k = g^{\text{feat}}(W^{\text{feat}}[c_i; c_j] + b^{\text{feat}})$$

- Each merge adds a binary auxiliary variable $y_k$
- A labeling $y$ of all auxiliary variables corresponds to a segmentation $s(y)$ of the image, provided that the label of a child is greater than or equal to that of its parent
- Model distribution over $y$ depends on $\psi$:

$$p(y \mid x, z; \theta) = \frac{1}{Z(x, z; \theta)} \exp \sum_{v \in \mathcal{N}(z)} \psi_v y_v$$




## LEARNING

- Minimize the KL divergence of $p$ from target distribution $q$

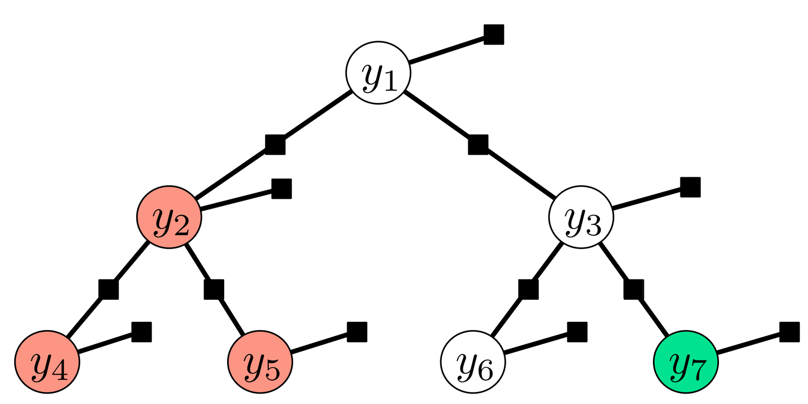
$$\mathcal{L}(x, z, S; \theta) = -\sum_{y \in \mathcal{Y}(z)} q(y|z, S) \log p(y|x, z; \theta) - H(q)$$

- $q$ is a mixture of distributions, one for each $s_j$:

$$q(y|z, S) = \frac{1}{M} \sum_{j=1}^{M} q_j(y|z, s_j), \text{ where}$$

$$q_j(y|z, s_j) = \frac{1}{Z_{q_j}(z, s_j)} \exp\left(-\frac{\Delta(s_j, s(y))}{\rho}\right)$$

- $p(y|x, z; \theta)$ can be encoded by a tree-structured factor graph with pairwise potentials encoding restrictions on $y$, leading to efficient and exact inference.
- $\Delta(s_j, s)$ decomposes over nodes and thus can also be represented by a factor graph with the same structure.
  - Inference is efficient for both $p$ and $q_j$
- Gradient updates for $\theta$ can be computed via back-propagation through structure (Goller & Kuchler 1996).
- With entropy message passing (Ilic et al. 2011), we can also compute a bound on the objective $\mathcal{L}(x, z, S; \theta)$.

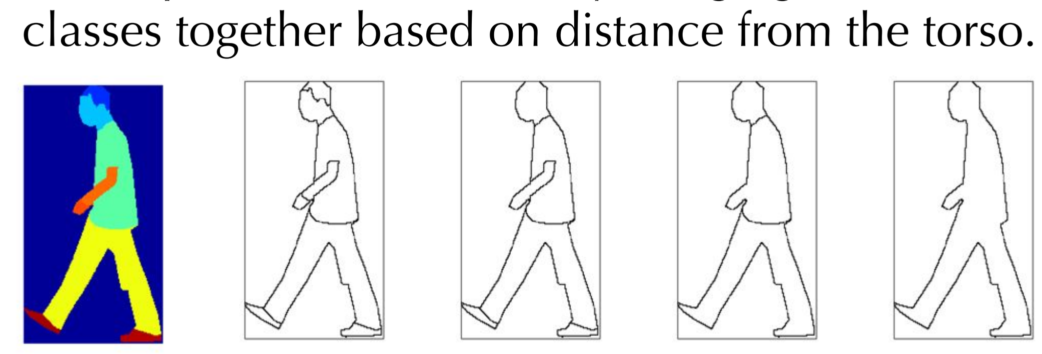



Factor graph structure for $p$ and $q_j$.
Example $y$ shown with all labels $y_i = 1$ highlighted.

The segmentation resulting from example labeling $y$. $y_6 = 0$ in this example, so $z_{12}$ (gold) and $z_{13}$ (blue) are separate in the corresponding segmentation.

## EXPERIMENTS

- Experiments will be conducted on segmentations synthesized from labeled body parts of the Penn-Fudan pedestrian dataset* by merging semantic classes together based on distance from the torso.



Semantic labels (left) and four synthesized ground truth segmentations

- Performance will be evaluated via precision (expected loss of output segmentations relative to closest ground truth) and recall (mean expected loss of output segmentations relative to each individual ground truth).
- Baselines will include alternate methods for generating multiple outputs, such as diverse M-best MAP (Batra et al. 2012).

*http://www.cis.upenn.edu/~jshi/ped_html/