Weakly Supervised Semantic Segmentation with Latent Conditional Random Fields

Gaurav Pandey

PhD advisor: Dr. Ambedkar Dukkipati

Department of Computer Science and Automation

gaurav.pandey@csa.iisc.ernet.in

Objective

Given a training set that comprises image and image-level labels only, infer the pixel-level labels of a test set that contains only images.

Other problems addressed during PhD

Generative Models

- Hierarchical completely random measures for topic modelling (ICML-16)
- A variational approach to deep conditional generative modelling (IJCNN-17)

Discriminative Models

- Deep neural networks of infinite width (ICML-14)
- Continuous learning with deep nonparametric neural networks (under progress)
- Discriminative Bayesian clustering (under progress)
- Latent conditional random fields for semantic segmentation (under submission)

The proposed model

1. We treat the pixel-level labels as the latent features of a CRF.
2. The pixels and the image-level label are the observed features.

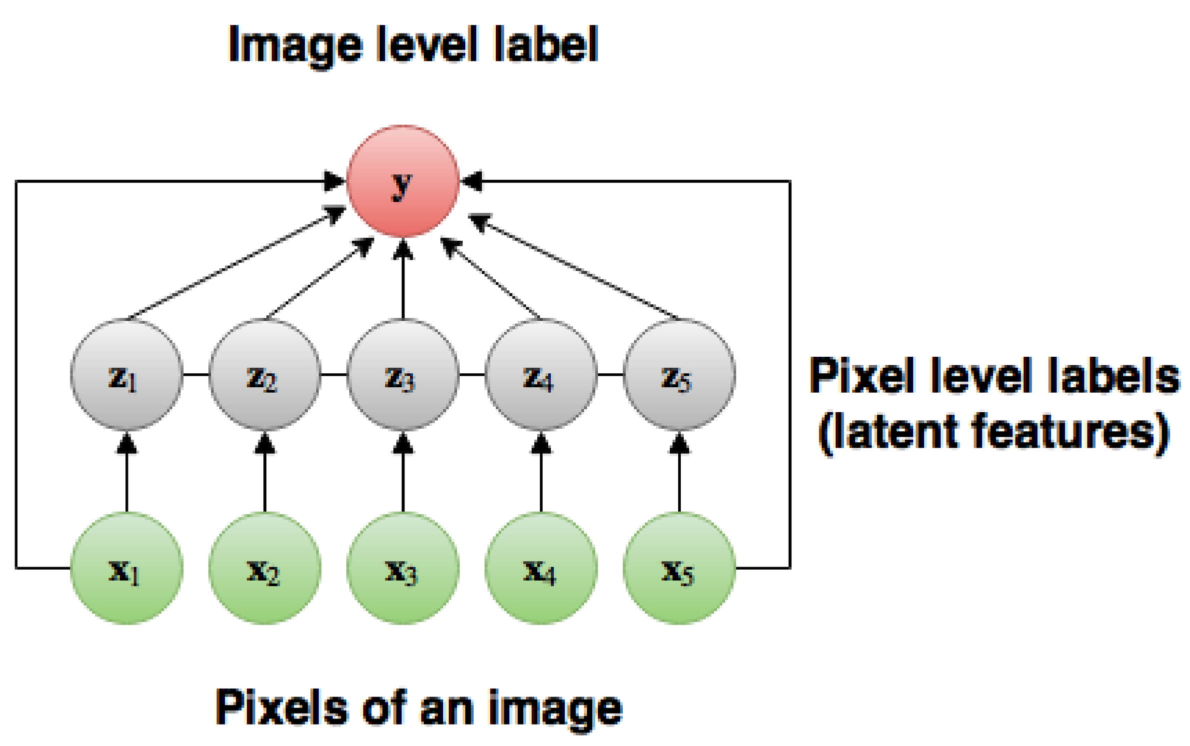


Figure 1: The latent CRF

3. $P(\mathbf{z}|\mathbf{x}) \propto \exp(-\sum_{j<i} k(\mathbf{x}_i, \mathbf{x}_j)\mu(\mathbf{z}_i, \mathbf{z}_j))$
4. Enforces neighboring pixels with similar color to also have the same label (local consistency constraint).
5. The aim is to maximize $P(\mathbf{y}|\mathbf{x})$.

$$P(\mathbf{y}|\mathbf{x}) = \sum_{\mathbf{z}} P(\mathbf{y}|\mathbf{z})P(\mathbf{z}|\mathbf{x})$$

6. Computation of $P(\mathbf{y}|\mathbf{x})$ is intractable.
7. Hence, we maximize its lower bound.

Variational lower bound

1. We chose a variational distribution $q(\mathbf{z}|\mathbf{x}, \mathbf{y})$, and obtain a lower bound on the log-likelihood.

$$\log p(\mathbf{y}|\mathbf{x}) \geq -\mathrm{KL}(q(\mathbf{z}|\mathbf{y}, \mathbf{x})||p(\mathbf{z}|\mathbf{x})) + \mathbb{E}_{q(\mathbf{z}|\mathbf{x},\mathbf{y})} \log p(\mathbf{y}|\mathbf{z}, \mathbf{x})$$

2. In this work, we assume that the variational distribution q factorizes completely, that is

$$q(\mathbf{z}|\mathbf{x}, \mathbf{y}) = \prod_{j=1}^{m} q(z_j|\mathbf{y}, \mathbf{x}) \tag{1}$$

3. Moreover,

$$q(z_{jk} = 1|\mathbf{x}, \mathbf{y}) = \frac{\exp(g_{jk}(\mathbf{x}))}{\sum_{k'=1}^{K} \exp(g_{jk'}(\mathbf{x}))} \equiv \varphi_{jk}(\mathbf{x}) , \tag{2}$$

where g is a fully convolutional neural network and $\{g_{jk}(\mathbf{x}), 1 \leq j \leq m, 1 \leq k \leq K\}$, are the outputs of g, when $\mathbf{x}$ is fed as input.

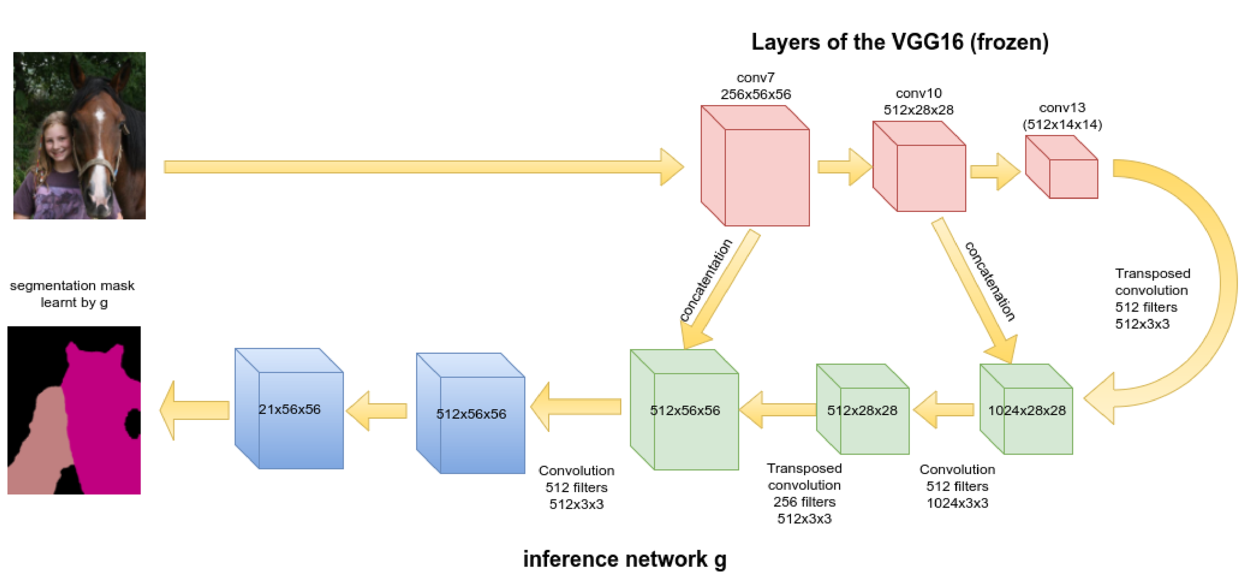


Figure 2: Inference network

4. The KL-divergence term forces the variational distribution to be close to the prior.
5. This ensures that the output of $q(\mathbf{z}|\mathbf{x}, \mathbf{y})$ also respects local consistency.
6. The second term ensures that pixel-level labels are consistent with the global labels.

Gradient of the lower bound

1. The first term is the KL-divergence loss on the output of CNN g, whose gradient can be computed exactly.
2. The gradient of the second term is approximated using reparametrization.
3. The multinomial distribution q is approximated by its continuous relaxation, the Gumbel-softmax distribution.
4. Next, samples from the Gumbel-softmax approximation are generated and fed to the classification network.
5. The gradient of $\log p(\mathbf{y}|\mathbf{x}, \mathbf{z})$ is computed with respect to the relaxed sample and backpropagated.

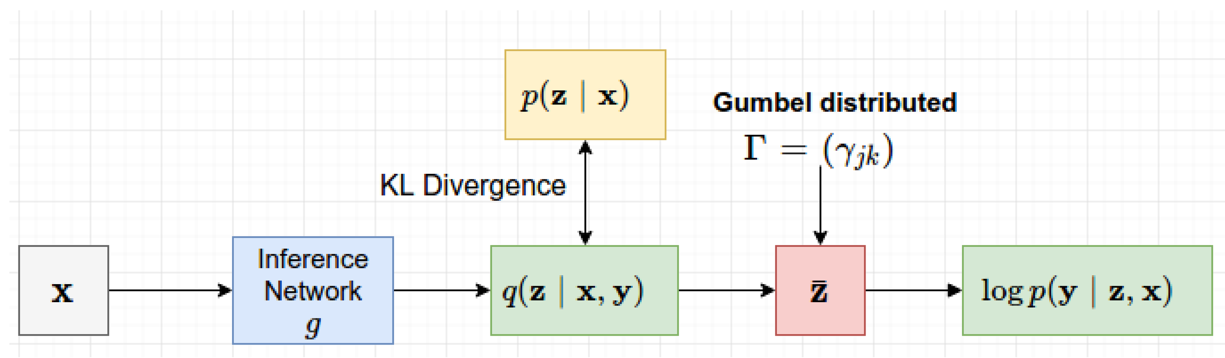


Figure 3: Implementation of the proposed model

6. Note that $\log p(\mathbf{y}|\mathbf{z}, \mathbf{x})$ contains no trainable parameters.

Qualitative Results on VOC 2012 dataset

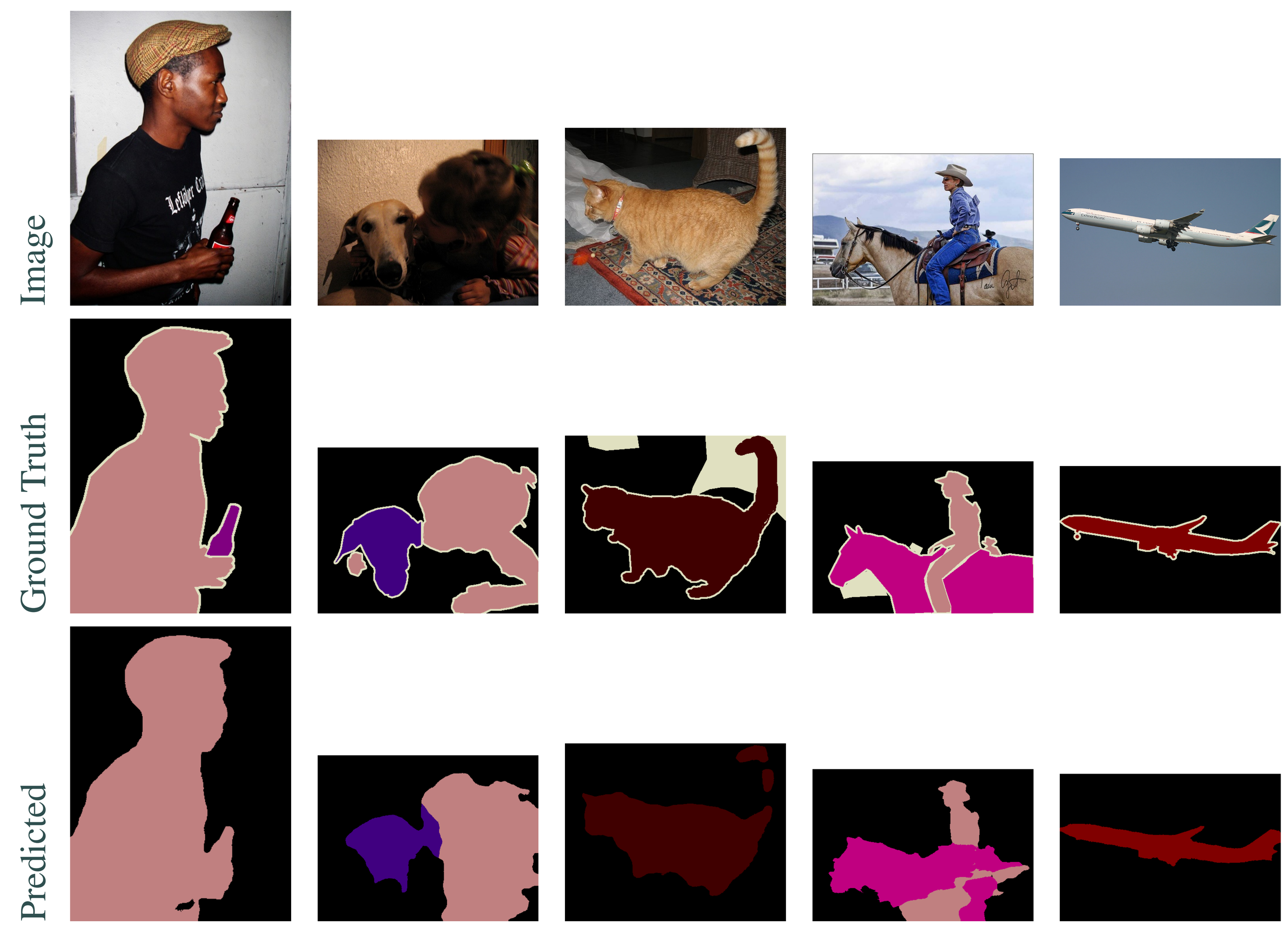


Table 1: Segmentation masks predicted by the model

Quantitative Results on VOC 2012 dataset

Saliency maps localize the important regions of an image, and can drastically improve performance.

1. Approaches that don't use saliency maps: (a) **MIL+ILP** (b) **EM-Adapt** (c) **CCNN**
2. Approaches that use saliency maps: (a) **SEC** (b) **STC**
3. Intersection over Union of predicted segmentation masks is given by:

$$IoU = \frac{\text{true positive}}{\text{true positive} + \text{false positive} + \text{false negative}}$$

class	MIL+ILP	EM-Adapt	CCNN	SEC	STC	Ours
background	77.2	67.2	68.5	82.4	84.5	**84.75**
aeroplane	37.3	29.2	25.5	62.9	68.0	**72.36**
bike	18.4	17.6	17.0	**26.4**	19.5	25.2
bird	25.4	28.6	25.4	61.6	60.5	**64.1**
boat	28.2	22.2	20.2	27.6	**42.5**	29.6
bottle	31.9	29.6	26.3	38.1	44.8	**53.6**
bus	41.6	47.0	46.8	66.6	**68.4**	53.1
car	48.1	44.0	47.1	62.7	**64.0**	62.9
cat	50.7	44.2	48.0	**75.2**	64.8	70.5
tvmonitor	35.0	31.6	36.9	45.3	31.2	**51.6**
mean	36.6	33.8	35.3	**50.7**	49.8	50.4

Table 2: IoU of predicted segmentation masks

Conclusions

1. This is the only work that uses a CNN as inference networks in a CRF for semantic segmentation.
2. The proposed model drastically outperforms all methods that don't use saliency maps.
3. The proposed model achieves performance comparable with other methods that use saliency maps.
4. The proposed models suggests that traditional probabilistic models, when combined with deep networks can achieve drastically improved performance.

Weakly Supervised Semantic Segmentation with Latent Conditional Random Fields

Gaurav Pandey

PhD advsior: Ambedkar Dukkipati
Department of Computer Science & Automation
Indian Institute of Science

Outline

Problems addressed during PhD

- Generative models
- Discriminative models

Semantic Segmentation

- Weakly Supervised Semantic Segmentation
- Conditional Random Fields
- Amortized Inference
- Experiments

Outline

Problems addressed during PhD

Generative Models

- Hierarchical completely random measures for topic modelling [Pandey and Dukkipati, 2016a] (ICML)
- A variational approach to deep conditional generative modelling [Pandey and Dukkipati, 2016b] (IJCNN)

Discriminative Models

- Deep neural networks of infinite width [Pandey and Dukkipati, 2014] (ICML)
- Continuous learning with deep nonparametric neural networks (under progress)
- Discriminative Bayesian clustering (under progress)
- Latent conditional random fields for semantic segmentation (under submission)

Problems addressed during PhD

Generative Models

- Hierarchical completely random measures for topic modelling [Pandey and Dukkipati, 2016a] (ICML)
- A variational approach to deep conditional generative modelling [Pandey and Dukkipati, 2016b] (IJCNN)

Discriminative Models

- Deep neural networks of infinite width [Pandey and Dukkipati, 2014] (ICML)
- Continuous learning with deep nonparametric neural networks (under progress)
- Discriminative Bayesian clustering (under progress)
- Latent conditional random fields for semantic segmentation (under submission)

Outline

- Problems addressed during PhD
  - Generative models
  - Discriminative models
- Semantic Segmentation
  - Weakly Supervised Semantic Segmentation
  - Conditional Random Fields
  - Amortized Inference
  - Experiments

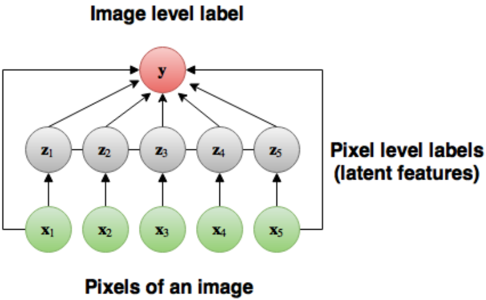
Objective

Weakly Supervised Semantic Segmentation

Given a training set that comprises image and image-level labels only, infer the pixel-level labels of a test set that contains only images.

Model

- We treat the pixel-level labels as the latent features of a conditional random field.
- The pixels and the image-level label are the observed features.



Conditional Random Field

- **Prior:** $P(\mathbf{z}|\mathbf{x}) \propto \exp(-\sum_{j<i} k(\mathbf{x}_i, \mathbf{x}_j)\mu(\mathbf{z}_i, \mathbf{z}_j))$
- Enforces neighboring pixels with similar color to also have the same label (local consistency constraint).
- The aim is to maximize $P(\mathbf{y}|\mathbf{x})$.

$$P(\mathbf{y}|\mathbf{x}) = \sum_{\mathbf{z}} P(\mathbf{y}|\mathbf{z})P(\mathbf{z}|\mathbf{x})$$

- Computation of $P(\mathbf{y}|\mathbf{x})$ is intractable.
- Hence, we maximize its lower bound.

Variational Lower Bound

- We chose a variational distribution $q(\mathbf{z}|\mathbf{x}, \mathbf{y})$, and obtain a lower bound on the log-likelihood.

$$\log p(\mathbf{y}|\mathbf{x}) \geq -\mathrm{KL}(q(\mathbf{z}|\mathbf{y}, \mathbf{x})||p(\mathbf{z}|\mathbf{x})) + \mathbb{E}_{q(\mathbf{z}|\mathbf{x},\mathbf{y})} \log p(\mathbf{y}|\mathbf{z}, \mathbf{x})$$

- We assume that the variational distribution q factorizes completely, that is

$$q(\mathbf{z}|\mathbf{x}, \mathbf{y}) = \prod_{j=1}^{m} q(z_j|\mathbf{y}, \mathbf{x}) \tag{1}$$

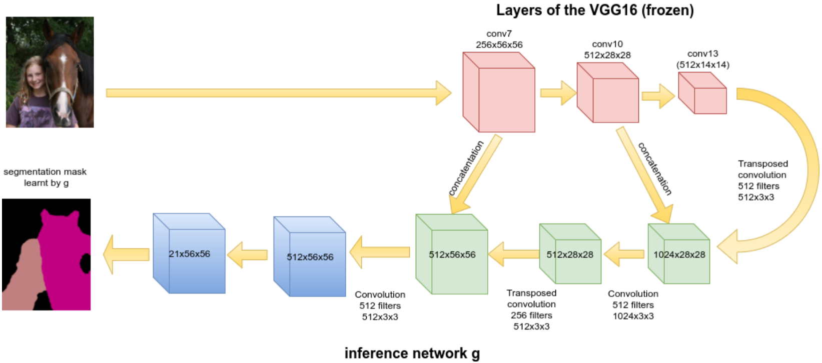
- Moreover,

$$q(z_{jk} = 1|\mathbf{x}, \mathbf{y}) = \frac{\exp(g_{jk}(\mathbf{x}))}{\sum_{k'=1}^{K} \exp(g_{jk'}(\mathbf{x}))} \tag{2}$$

where g is a fully convolutional neural network.

Variational Lower Bound

- The distribution $q(\mathbf{z}|\mathbf{x}, \mathbf{y})$ is parametrized by a CNN.

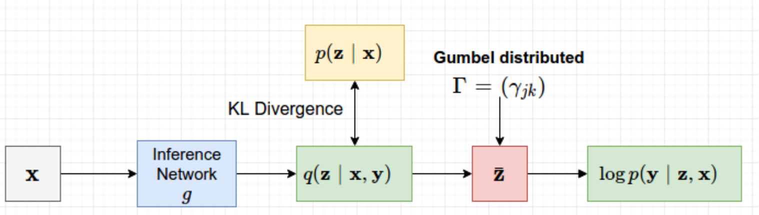


Variational Lower Bound

- The KL-divergence term forces the variational distribution to be close to the prior.
- This ensures that the output of $q(\mathbf{z}|\mathbf{x}, \mathbf{y})$ also respects local consistency.
- The second term ensures that pixel-level labels are consistent with the global labels.

Gradient of the Lower Bound

- The first term is the KL-divergence loss on the output of CNN g.
- The gradient of this loss can be computed exactly.
- The gradient of the second term is approximated using MCMC samples.



Qualitative results on VOC 2012 [1] dataset

Table: Examples of predicted segmentation masks. The middle row is the ground truth.

[1] [Everingham et al.,]

Compared methods

Saliency maps localize the important regions of an image, and can drastically improve performance.

- Approaches that don't use saliency maps:
  - **MIL+ILP [Pinheiro and Collobert, 2015]**
  - **EM-Adapt [Papandreou et al., 2015]**
  - **CCNN [Pathak et al., 2015]**
- Approaches that use saliency maps:
  - **SEC [Kolesnikov and Lampert, 2016]**
  - **STC [Wei et al., 2016]**

Evaluation metric

- For each pixel, the class label is predicted.
- For each class, intersection over union (IoU) score is calculated as:

$$\frac{\text{true positive}}{\text{true positive} + \text{false positive} + \text{false negative}}$$

- The mean IoU is simply the average over all the classes.

Quantitative results on VOC 2012 dataset

class	MIL+ILP	EM-Adapt	CCNN	SEC	STC	Ours
background	77.2	67.2	68.5	82.4	84.5	**84.75**
aeroplane	37.3	29.2	25.5	62.9	68.0	**72.36**
bike	18.4	17.6	17.0	**26.4**	19.5	25.2
bird	25.4	28.6	25.4	61.6	60.5	**64.1**
boat	28.2	22.2	20.2	27.6	**42.5**	29.6
bottle	31.9	29.6	26.3	38.1	44.8	**53.6**
bus	41.6	47.0	46.8	66.6	**68.4**	53.1
car	48.1	44.0	47.1	62.7	**64.0**	62.9
mean	36.6	33.8	35.3	**50.7**	49.8	50.4

Table: Results on PASCAL VOC 2012 (IoU in %) *val* set.

Conclusions

- The first work that uses a CNN as inference networks in a CRF for semantic segmentation.
- The proposed model drastically outperforms all methods that don't use saliency maps.
- The proposed model achieves performance comparable with other methods that use saliency maps.
- The proposed models suggests that traditional probabilistic models, when combined with deep networks can achieve drastically improved performance.

References I

Everingham, M., Van Gool, L., Williams, C. K. I., Winn, J., and Zisserman, A.
The PASCAL Visual Object Classes Challenge 2012 (VOC2012) Results.
http://www.pascal-network.org/challenges/VOC/voc2012/workshop/index.html.

Kolesnikov, A. and Lampert, C. H. (2016).
Seed, expand and constrain: Three principles for weakly-supervised image segmentation.
In *European Conference on Computer Vision*, pages 695–711. Springer.

References II

Pandey, G. and Dukkipati, A. (2014).
Learning by stretching deep networks.
In *Proceedings of The 31st International Conference on Machine Learning*, pages 1719–1727.

Pandey, G. and Dukkipati, A. (2016a).
On collapsed representation of hierarchical completely random measures.
In *Proceedings of The 33rd International Conference on Machine Learning*, pages 1605–1613.

Pandey, G. and Dukkipati, A. (2016b).
Variational methods for conditional multimodal deep learning.
arXiv preprint arXiv:1603.01801.

References III

Papandreou, G., Chen, L.-C., Murphy, K., and Yuille, A. L. (2015).
Weakly-and semi-supervised learning of a dcnn for semantic image segmentation.
arXiv preprint arXiv:1502.02734.

Pathak, D., Krahenbuhl, P., and Darrell, T. (2015).
Constrained convolutional neural networks for weakly supervised segmentation.
In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1796–1804.

References IV

Pinheiro, P. O. and Collobert, R. (2015).
From image-level to pixel-level labeling with convolutional networks.
In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1713–1721.

Wei, Y., Liang, X., Chen, Y., Shen, X., Cheng, M.-M., Feng, J., Zhao, Y., and Yan, S. (2016).
Stc: A simple to complex framework for weakly-supervised semantic segmentation.
IEEE Transactions on Pattern Analysis and Machine Intelligence.