# Topology Aware Fully Convolutional Networks For Histology Gland Segmentation

Aïcha BenTaieb and Ghassan Hamarneh

Medical Image Analysis Lab, Simon Fraser University, Burnaby, Canada

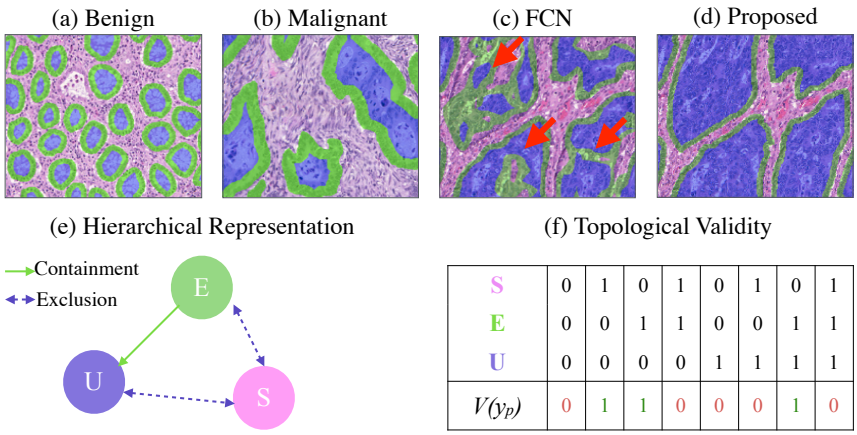
**Abstract.** The recent success of deep learning techniques in classification and object detection tasks has been leveraged for segmentation tasks. However, a weakness of these deep segmentation models is their limited ability to encode high level shape priors, such as smoothness and preservation of complex interactions between object regions, which can result in implausible segmentations. In this work, by formulating and optimizing a new loss, we introduce the first deep network trained to encode geometric and topological priors of containment and detachment. Our results on the segmentation of histology glands from a dataset of 165 images demonstrate the advantage of our novel loss terms and show how our topology aware architecture outperforms competing methods by up to 10% in both pixel-level accuracy and object-level Dice.

## 1 Introduction

Object segmentation, assigning semantic labels to pixels within an object, is a fundamental problem in medical image analysis. Reproducible classification or grading of adenocarcinomas benefits from accurate segmentation of epithelial glands from histology images [4, 6, 10]. Despite great advances in histology gland segmentation, many challenges remain. The complexity of glandular objects' appearance, which correlates with the degree of cancer differentiation (e.g. high grade tumours present degenerated glands), and the high variability in histology image acquisition (i.e. microscope, lighting, and staining) accounts for two of the major challenges in histopathology gland segmentation [15].

Generally, state-of-the-art segmentation techniques benefit from incorporating prior knowledge about the target structures into the segmentation formulation [2, 12]. Recent gland segmentation methods, e.g. [5, 15], are no exception as they do encode gland geometrical priors into their formulation, namely that glands are smooth tubular structures, composed of a central area (lumen) surrounded by epithelial cells forming a nuclear boundary around the lumen (examples in figure 1-(a,b)). However, a limitation of these works is that they rely on hand-crafted features (often pixel-level color and texture cues) to detect each glandular component, which can be susceptible to biological and staining variation. To counteract these problems, existing works commonly resort to ad-hoc post-processing methods for false negative removal and object delineation.

The recent success of deep convolutional networks (CNN) for object recognition and classification tasks has been leveraged for segmentation, or pixel-level




| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| S | 0 | 1 | 0 | 1 | 0 | 1 | 0 | 1 |
| E | 0 | 0 | 1 | 1 | 0 | 0 | 1 | 1 |
| U | 0 | 0 | 0 | 0 | 1 | 1 | 1 | 1 |
| $V(y_p)$ | 0 | 1 | 1 | 0 | 0 | 0 | 1 | 0 |

**Fig. 1:** Multi-region gland representation. Example delineations of (a) benign and (b) malignant colon adenocarcinoma glands. Example of segmentation output of (c) FCN trained without topological priors and of (d) FCN trained with a topology-aware loss. Topological violations are indicated with red arrows. (e) The topological relationships between the multi-region gland components. (f) Topological validity indicator $V(y_p)$ for each possible labelling $y$ of a pixel $p$. Blue regions in (a,b) represent the inside glandular lumen as well as goblet cells if present (denoted U in (e,f)). Green regions delineate the Epithelial boundary around the gland (E in (e,f)). The background (purple) indicate stromal nuclei (S in (e,f)).

classification, through the introduction of fully convolutional networks (FCN) [8], in which all fully-connected layers of a standard classification CNN are converted into convolutional layers. FCNs have been proven capable of learning high-level complex hierarchies of descriptive and discriminative features useful for per-pixel predictions [8, 9, 13]. Models inspired by FCN architectures were successfully applied and adapted to various biomedical image segmentation applications [13].

Despite their success, FCN-based segmentations suffer from relying on a pixel-level prediction that is not designed to account for higher-order properties, such as boundary smoothness and the topological label interactions of multi-part objects (as in the lumen and epithelium of glands). Moreover, FCNs tend to produce low-resolution segmentations due to the subsampling resulting from stacked layers of convolutions and pooling. To overcome FCN's limitations, different strategies have been explored to preserve object boundaries. One approach consists of adding trainable upsampling layers using deconvolution operations [8, 9]. While these layers are useful in reconstructing the input image size from coarser outputs, they only partially recover object boundaries. Other approaches attach a dense conditional random field (CRF) to the FCN, either as a post-processing step [1] or jointly trained with the FCN [16], in order to increase the sharpness of the output. However, both approaches require extra computational costs for optimizing the CRF and only specific graphical models can be integrated into the FCN learning pipeline. To the best of our knowledge, none of the existing works incorporates topology priors in the learning of FCNs.

In this work, we propose to encode smoothness of, and topological constraints between, segmented regions of spatially-recurring, multi-part objects (e.g. several glands, each with lumen and epithelium) into the learning of FCNs. Our aim is to train a deep network that produces topologically plausible, high-resolution segmentation output. Our strategy is to design a loss function with specific penalty terms that encode the desired boundary smoothness priors and hierarchical relationships between regions labels. In our specific application, the multi-region relations correspond to containment and exclusion properties observed between the smooth lumen and epithelial gland boundaries (figure 1-c,d,e).

Our proposed loss exploits the elegant graph formulation of hierarchical label relationships used in the context of image classification [3], and the popular energy-based multi-region labelling framework introduced by Delong and Boykov [2]. In contrast to these previous works, our formulation is specifically designed for object segmentation and pixel-level interactions in an end-to-end trainable deep network. Further, our formulation does not require post-hoc processing or additional heavy, test-time computational costs associated with the previously explored CRF optimization based approaches. Extensive experiments on the publicly available Warwick-QU dataset of histology colon glands and on different FCN architectures and training strategies (e.g. combining FCNs with CRFs) demonstrate the advantage of our method in learning more regularized deep networks for gland segmentation.

## 2 Method

Our goal is to incorporate topological priors: containment and exclusion, and geometrical prior: boundary smoothness, into the learning of deep fully convolutional networks. In the context of histology glands, there is a containment relation between lumen and epithelial boundary and an exclusion relation between stroma and all other regions (figure 1-(e)). We also know that a smooth epithelial boundary separates the lumen from the stroma (figure 1-(a,b)).

We train an FCN from a set of images and their corresponding ground truth segmentations, $\{(x^{(n)}, y^{(n)}); n = 1, 2, \ldots, N\}$. We drop the superscript $(n)$ when referring to any image $x$ or segmentation $y$. The FCN's prediction of $y$ is denoted $y^*$. A (crisp) segmentation of a color image $x \in \mathcal{R}^{H \times W \times 3}$ assigns the $p$-th pixel $x_p$ in $x$ a vector $y_p = (y_p^1, y_p^2, ..., y_p^L) \in \{0, 1\}^L$, where $y_p^r$ indicates whether pixel $x_p$ belongs to region $r$, and $L$ is the number of region labels.

**FCN's per-pixel loss:** Training an FCN for segmentation amounts to finding the network's parameters $\theta$ that solves the following optimization:

$$\theta^* = \arg\min_{\theta} \sum_{n=1}^{N} \mathcal{L}(x^{(n)}; \theta), \tag{1}$$

$$\mathcal{L}(x; \theta) = \sum_{p \in \Omega} \sum_{r=1}^{L} -y_p^r \log P(y_p^r = 1 | x_p; \theta), \;\; P(y_p^r = 1 | x_p; \theta) = \frac{\exp(a_r(x_p))}{\sum_{k=1}^{L} \exp(a_k(x_p)} \tag{2}$$

where $\Omega$ is the pixel space, $\mathcal{L}$ is the multinomial cross-entropy loss, and $P$ are the class probabilities output of the softmax function of the FCN, which is based on $a_r(x_p)$, the output activation for region $r$ and pixel $p$. $\mathcal{L}$ measures the compatibility between the predictions $P(y_p^r = 1)$ and the corresponding ground truth $y_p^r$ for each pixel $x_p$ in the training dataset.

**Multi-region interactions:** We now modify (1) and (2), by introducing additional hierarchical relations between region labels and add a regularization term, and perform the following minimization of the new topology-aware loss:

$$\theta^* = \arg\min_{\theta} \sum_{n=1}^{N} \alpha_1 \mathcal{L}_T(x^{(n)}; \theta) + \alpha_2 \mathcal{L}_S(x^{(n)}; \theta); \tag{3}$$

where $\mathcal{L}_T$ and $\mathcal{L}_S$ refer, in order, to the pixel-level loss functions that encode the topological relations between labels and the smoothness constraints. We elaborate on the design of each term of the proposed loss below. Note that $\alpha_1$ and $\alpha_2$ are user-defined weights used to balance the contribution of each prior. We discuss the impact of these terms in the Experiments.

**Hierarchical label relations:** The goal here is to define $\mathcal{L}_T$ such that the network is trained, not only to penalize incorrect label assignment per pixel, but to also penalize incorrect label hierarchy. In gland segmentation, for example, the fact that region U (lumen) should be contained in region E (epithelium), not only requires $P(y_p^U = 1)$ to be high at a lumen pixel $p$ but so should $P(y_p^E = 1)$ and $P(y_p^S = 0)$. In other words, the joint probability $P(y_p^S = 0, y_p^U = 1, y_p^E = 1)$ should be high. Given $L$ labels or tissue classes, there are $2^L$ possible assignments per pixel (figure 1-f). Some of these assignments would be plausible, as they respect the label hierarchy imparted by the containment and exclusion priors, while others would not. Inspired by the strategy used in [3], which introduces a generic CRF-based approach for image classification with structured label relations, we define the following unary loss:

$$P(y_p|x_p;\theta) = \frac{1}{Z} \prod_{r=1}^{L} \exp\big(a_r(x_p)\big) \times y_p^r \times V(y_p), \quad Z = \sum_{r=1}^{L} P(y_p^r|x_p;\theta); \tag{4}$$

where $P$ is the normalized joint probability for the label vector $y_p$, $Z$ is the partition function, $a_r(x)$ is the FCN's output prediction for region label $r$ and $V(y_p) \in \{0, 1\}$ is a validity indicator function returning 1 if a given label vector $y_p$ corresponds to a topologically-valid assignment, and zero otherwise (see figure 1-(f)). The probability of a region $r$ is computed by marginalizing all other region labels: $P(y_p^r = 1|x_p;\theta) = \sum_{y_p : y_p^r = 1} P(y_p|x_p;\theta)$.

Combined with a softmax loss, the hierarchical probabilities $P(y_p^r|x_p;\theta)$ form our first penalty term $\mathcal{L}_T$. Note that if all regions are mutually exclusive, $P(y_p^r = 1|x_p;\theta)$ is equivalent to the softmax probability defined in (2).

**Pairwise penalties:** The goal here is to define $\mathcal{L}_S$ such that the network is trained to produce segmentations with smooth boundaries. We encode this geometrical property via a binary pairwise label interaction softmax loss:

$$\mathcal{L}_S(x;\theta) = \sum_{p\in\Omega}\sum_{r=1}^{L}\sum_{q\in\mathcal{N}^p} B_{p,q} \times y_p^r \left| P(y_p^r|x_p;\theta) - P(y_q^r|x_q;\theta) \right| ; B_{p,q} = \begin{cases} 1 \text{ if } y_p^r = y_q^r \\ 0 \text{ else} \end{cases} \tag{5}$$

where $\mathcal{N}^p$ corresponds to the 4-connected neighborhood of pixel $p$. $\mathcal{L}_S$ trains the network to output *regularized* pairs of softmax label probabilities of neighbouring pixels $p$ and $q$ (i.e. having similar predicted probabilities) when ground truth pixel pairs belong to the same tissue label ($B_{p,q} = 1$). At the same time, $\mathcal{L}_S$ trains the network to allow discontinuities across tissue boundaries ($B_{p,q} = 1$).

**Optimization and inference:** The proposed loss is optimized using stochastic gradient descent. To infer the output predictions $y^*$ (e.g. a probability score for each region and each pixel), a simple forward pass through the trained network is required. Probabilities are computed following the label relations defined in (4). The final binary output segmentation $y^*$ corresponds to the region with maximum probability per pixel.

## 3 Experiments

The **implementation** of our proposed model was realized as a new loss layer in Caffe deep learning library [7] and can be used on top of any fully convolutional model given multi-region relations. Given the large input image size ($500 \times 500$), we used a mini-batch size of 1 with a momentum of 0.99. The learning rate was tuned for each model on a validation set during training. We used the totality of the publicly available Warwick-QU colon adenocarcinoma dataset released as part of the GlaS Challenge [14], which consists of 85 training and 80 test images. In all experiments, we used 70 images for training, 15 for validation and 80 for test. We kept the training and test splits provided by the challenge organizers. We used a series of elastic (warping) and affine transformations (rotation, scaling, color shifts) to augment the training dataset by a factor of $\sim 150$. All models were trained on a NVIDIA 12 GB GPU card and training time ranged between 2 hours for relatively small models ($\sim$6 layers) and 36 hours for deeper models. Test times were $\sim$1 s/image for all models.

To test the **advantage of adding topological priors** in the learning of FCN, we compared the performance of four network architectures that implement different sampling strategies for border sharpening and with increasing number of layers, trained with vs. without including our multi-region priors $\mathcal{L}_T$ and $\mathcal{L}_S$: i) Alexnet-FCN and ii) FCN-8s (with a stride of 8) [8] use a simple bilinear interpolation for upsampling; iii) U-Net [13] includes bridge-like layers between coarser layers' outputs and finer layers; whereas iv) DN [9] uses deconvolution layers as upsampling strategy.

We used two evaluation metrics: 1) pixel-level accuracy and 2) object-level Dice similarity coefficient (figure 2). Our results show that, for the same optimizer and the same network complexity, using our proposed loss yields an average improvement of 9 to 15% in correctly labelling pixels and 3 to 5% in delineating glands.

We tested the **robustness of our results** to the hyper-parameters in (3). We used a validation set to tune these parameters and found that regardless of the model's architecture using equally weighted penalty terms generally gave us best results. We

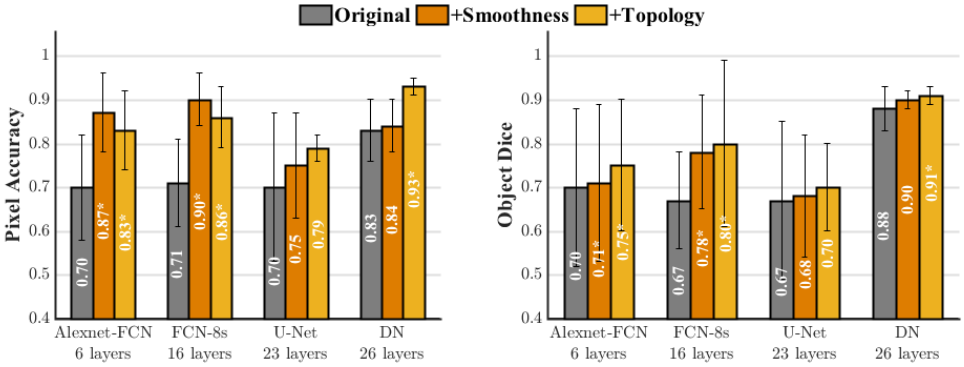



**Fig. 2:** Advantage of the proposed loss: "Origina" refers to the cross-entropy loss $\mathcal{L}$. "+Smoothness" refers to using the proposed penalty term $\mathcal{L}_S$ and "+Topology" refers to adding our topology prior $\mathcal{L}_T$. The asterisk (*) corresponds to statistically significant differences from the original models obtained using a Wilcoxon matched-pairs signed rank sum test at $p<0.05$.

observed a minimal change in pixel accuracy and object Dice (less than $1e^{-4}$) when varying the difference between $\alpha_1$ and $\alpha_2$ by $\pm 20\%$.

We also compared our method with the winner of the GlaS Challenge [14], CuMedVision2, which also used a FCN-based model with a special upsampling strategy. Note that winners' model architecture was not released and only the number of pooling layers were reported [14]. For fair comparison we report results with FCN-8s that has similar number of pooling layers. Using our topology-aware loss with FCN-8s architecture, we outperformed the reported results of CuMedVision2 by 18% for F1 score, 3% for object Dice but CuMedVision2 surpassed our approach by 12% in terms of Hausforff distance.

To compare applying the proposed **loss penalties vs. graphical models**, we test the performance of FCN-32s (with a stride of 32) trained with $\mathcal{L}_T + \mathcal{L}_S$ with: a) the original FCN-32s model that optimizes per-pixel loss ($\mathcal{L}$), and with two methods that refine FCN's segmentation by incorporating a probabilistic graphical model optimization: b) DeepLab [1], which uses a special fully-connected CRF, where the pairwise terms depend on pixels positions and color intensities as a post-processing step, and c) CFR-RNN [16], where the same CRF model is jointly trained with the FCN. In

| Method | Pixel Accuracy | Object Dice | Inference |
|---|---|---|---|
| FCN-32s [8] | 0.80 ± 0.12 | 0.70 ± 0.17 | **28.62s** |
| DeepLab [1] | 0.78 ± 0.12 | 0.69 ± 0.19 | 38.02s |
| CRF-RNN [16] | 0.73 ± 0.19 | 0.42 ± 0.12 | 32.50s |
| FCN+Smoothness | **0.86 ± 0.07** | 0.78 ± 0.11 | **28.62s** |
| FCN+Smoothness+Topology | 0.76 ± 0.09 | **0.80 ± 0.12** | 28.63s |

**Table 1:** Penalty terms vs. graphical models. +Smoothness refers to adding $\mathcal{L}_S$ in the FCN-32s training. +Smoothness+Topology refers to $\mathcal{L}_T + \mathcal{L}_S$.

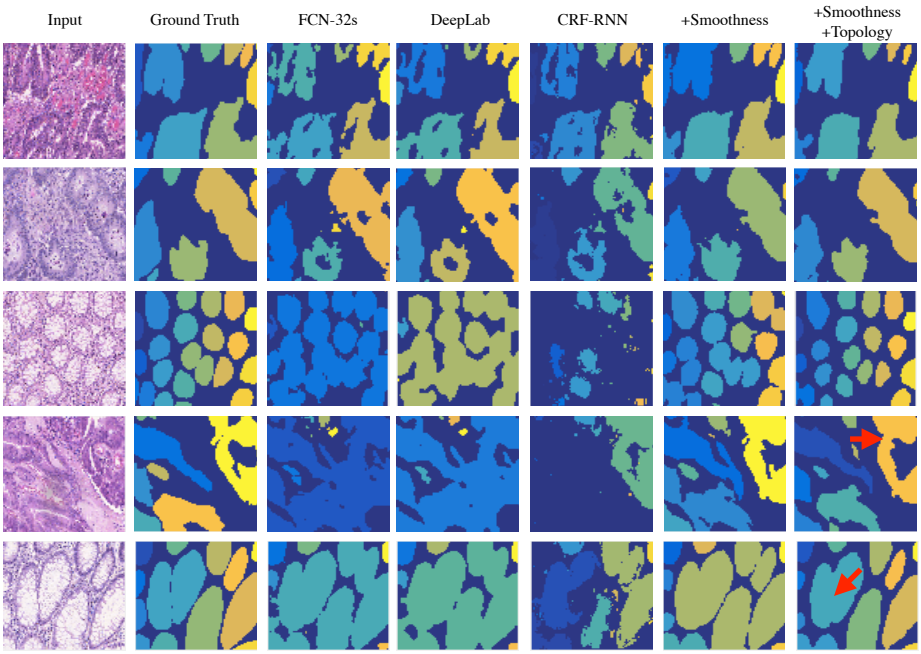



**Fig. 3:** Qualitative comparisons. Note the smoother boundaries and individually detected glands produced by our method (last two columns). Red arrows highlight challenging cases that were not successfully segmented.

DeepLab and CRF-RNN, the CRF energy function is optimized using iterations of the mean field approximation.

As shown in table 1, using our additional smoothness and topology priors in the training of FCN-32s, our model achieves 13 to 38% higher object Dice compared to the original FCN-32s, DeepLab or CRF-RNN. It is also worth pointing out that our proposed method does not incur any additional computational cost during inference, contrarily to DeepLab and CRF-RNN.

It is worth noting that DeepLab and CRF-RNN degrade the performance of the original FCN-32s model. This initially surprising result may be explained by the fact that the special CRF model used in DeepLab and CRF-RNN includes image (color)-based pairwise terms in their energy functions, which are sensitive to stain variations among glands and between stroma and glands.

Finally, we observe that adding our topology priors result in an increase in Dice by 10% over FCN-32s (0.70% to 0.80%) despite a smaller decrease of 4% in pixel accuracy (0.80% to 0.76%). This implies that the additional priors are critical for the detection of *individual* glands, particularly due to how the topology prior encodes relevant object-level (i.e. beyond pixel-level) information during training. Qualitative results are presented in figure 3. Adding topology penalties generally resulted in smoother boundaries and individually segmented glands. However, it did not fully compensate for the loss of fine-grained details resulting from upsampling the probabilities in some very challenging cases where glands' boundaries are extremely thin.

## 4 Conclusion

We hypothesized that the inclusion of prior knowledge in the training of deep fully convolutional networks for the segmentation of histology glands can result in more accurate segmentations. To test our hypothesis, we presented a novel loss function inspired by energy-based models for multi-region labelling and adapted for deep networks. Our findings show that our approach yields significantly more accurate and plausible segmentations while being more computationally efficient at test-time. We plan to further investigate the effect of equipping deep learning models with relevant prior knowledge for training more regularized networks on different medical segmentation applications.

**Acknowledgments.** We gratefully acknowledge NSERC for funding and NVIDIA Corporation for GPU donation.

## References

1. LC Chen et al. Semantic image segmentation with deep convolutional nets and fully connected CRFs. *arXiv:1412.7062*, 2014.
2. A Delong and Y Boykov. Globally optimal segmentation of multi-region objects. In *ICCV*, pages 285–292, 2009.
3. J Deng et al. Large-scale object classification using label relation graphs. In *ECCV*, pages 48–64. 2014.
4. CW Elston and IO Ellis. Pathological prognostic factors in breast cancer. the value of histological grade in breast cancer: experience from a large study with long-term follow-up. *Histopathology*, 19(5):403–410, 1991.
5. C Gunduz-Demir, M Kandemir, A B Tosun, and C Sokmensuer. Automatic segmentation of colon glands using object-graphs. *MedIA*, 14(1):1–12, 2010.
6. PA Humphrey. Gleason grading and prognostic factors in carcinoma of the prostate. *Modern pathology*, 17(3):292–306, 2004.
7. Y Jia et al. Caffe: Convolutional architecture for fast feature embedding. In *ACM Multimedia*, pages 675–678, 2014.
8. J Long, E Shelhamer, and T Darrell. Fully convolutional networks for semantic segmentation. In *CVPR*, pages 3431–3440, 2015.
9. H Noh, S Hong, and B Han. Learning deconvolution network for semantic segmentation. In *CVPR*, pages 1520–1528, 2015.
10. M S Nosrati, S Andrews, and G Hamarneh. Bounded labeling function for global segmentation of multi-part objects with geometric constraints. In *IEEE ICCV*, pages 2032–2039, 2013.
11. M S Nosrati and G Hamarneh. Local optimization based segmentation of spatially-recurring, multi-region objects with part configuration constraints. *TMI*, 33(9):1845–1859, 2014.
12. O Ronneberger, P Fischer, and T Brox. U-net: Convolutional networks for biomedical image segmentation. In *MICCAI*, pages 234–241. 2015.
13. K Sirinukunwattana et al. Gland segmentation in colon histology images: The GlaS challenge contest. *arXiv:1603.00275*, 2016.
14. K Sirinukunwattana, D Snead, and NM Rajpoot. A stochastic polygons model for glandular structures in colon histology images. *TMI*, 34(11):2366–2378, 2015.
15. S Zheng et al. Conditional random fields as recurrent neural networks. In *ICCV*, pages 1529–1537, 2015.