Learning Semi-supervised Multi-Label Fully Convolutional Network for Hierarchical Object Parsing

Xiaobai Liu, Qian Xu, Eric Medwedeff, Grayson Adkins, Liang Lin, Shuicheng Yan

***Abstract*—This paper presents a semi-supervised multi-label Fully Convolutional Network (FCN) for hierarchical object parsing of images. We consider each object part (e.g., eye and head) as a class label and learn to assign every image pixel to multiple coherent part labels. Different from previous methods that consider part labels as independent classes, our method explicitly models the internal relationships between object parts, e.g., that a pixel highly scored for eyes should be highly scored for heads as well. Such relationships directly reflect the structure of the semantic space and thus should be respected while learning the deep representation. We achieve this objective by introducing a multi-label softmax loss function over both labeled and unlabeled images and regularizing it with two pair-wise ranking constraints. The first constraint is based on a manifold assumption that image pixels being visually and spatially close to each other should be collaboratively classified as the same part label. The other constraint is used to enforce that no pixel receive significant scores from more than one labels that are semantically conflicting with each other. The proposed loss function is differentiable with respect to network parameters and hence can be optimized by standard stochastic gradient methods. We evaluate the proposed method on two public image datasets for hierarchical object parsing and compare it to the alternative parsing methods. Extensive comparisons showed that our method can achieve state-of-the-art performance while using 50% less labeled training samples than the alternatives .**

***Index Terms*—fully convolutional network, semi-supervised learning, hierarchical models**

I. Introduction

The goal of this work is to develop an effective approach capable of segmenting objects, object parts (e.g. head, torso, legs) and subparts (e.g. eyes, mouses) in images and generating a hierarchical representation of objects. The outcomes of our approach include a pixel-wise binary mask for each entity of the hierarchy, which can be used to assist in high-level image tasks, e.g., human pose recognition [15] or human interaction recognition [1] [33]. In the past decade, the state of object parsing has been rapidly evolving [39] [18] [13][14] [37], largely driven by the advances in statistical learning and computer vision. In particular, the recently developed Fully Convolutional Network (FCN) [27] is capable of end-to-end learning multi-level feature representations for semantic image segmentation. Multiple object parsing methods [31] [43] [9] utilize FCN as basic networks and achieved encouraging results on multiple object detection benchmarks. The learning of such deep representations, however, requires tens of thousands of labeled samples and hence requires cost-intensive human efforts to prepare training data. The situation becomes worse while dealing with hierarchical object parsing, where an image includes tens of part labels. Thus, there is a demand for developing weakly supervised deep models for object parsing in practical deployment.

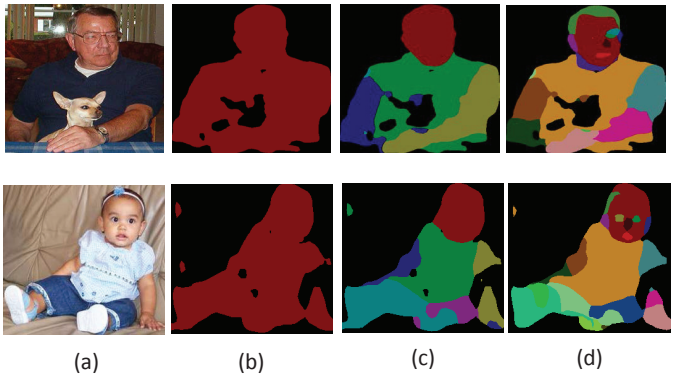

Fig. 1. Hierarchical object parsing. (a) Input images; (b)-(c): object masks segmented by the proposed method for 2 labels (human and non-human), 6 human part labels, and 31 human subpart labels, respectively.

Figure 1 shows two exemplar results of the proposed method for hierarchical object parsing. Given a single image as the input, our method can segment human region, human parts (e.g. head), and human subparts (e.g. eyes), as shown in the subfigures (b) - (d), respectively. These part labels are not semantically independent with each other during inference because, for example, a region of noses should be labeled as heads as well, and a region of hands is part of the region of upper-body. In addition to such coherences, two human part labels might be exclusive from each other. For example, a region cannot be classified as heads and torso simultaneously. An effective hierarchical object parsing algorithm has to respect both coherent and exclusive constraints between image labels, which has not been systemically studied in the literature

Xiaobai Liu is with the Department of Computer Science, San Diego State University, San Diego, CA 92150 USA. Email: xiaobai.liu@mail.sdsu.edu

Qian Xu is with the Xrelab Inc., San Diego, CA and Department of Computer Science, San Diego State University, San Diego, CA 92150 USA.

Eric Medwedeff is with the Department of Computer Science and Computational Science Research Center (CSRC), San Diego State University, San Diego, CA 92150 USA.

Grayson Adkins was with the Department of Computer Science, San Diego State University, San Diego, CA 92150 USA.

Liang Lin is with the Human Cyber Physical Intelligence Integration Lab, Sun Yat-Sen University, Guangzhou, China

Shuicheng Yan is with the 360/Qihu Inc., China and the National University of Singapore, Singapore.

of deep representation learning [4].

In the proposed method, we consider each object part of the hierarchy as a class label and aim to learn a multi-label FCN from images. Hierarchical object parsing is essentially a multi-label image segmentation problem which aims to assign each pixel of the input image to multiple part labels, e.g., eye, head, and upper-body. While the conventional neural network methods [27] can be used for multi-label settings, their loss functions cast image labels to be semantically independent with each other and ignore their coherent relationships. For example, an image pixel receiving the label of eye should be classified as a head as well but not vice versa. In this work, we propose to develop a multi-label softmax loss for FCN to encourage compatibilities between the predicated labels and ground-truth labels while respecting the above coherence constraints between image labels.

We will learn the proposed model using a small number of annotated images and a large amount of raw images without annotations. This semi-supervised setting allows our model to generalize to unseen data samples and avoid potential over-fitting with the training data. Over-fitting is a serious issue for modern neural network techniques [4], which have been employing an increasingly large set of network parameters (e.g., with deeper layers or more hidden units). To suppress the effects of over-fitting, in this work, we develop two regularization terms for the proposed semi-supervised model.

- **Manifold regularization**. We introduce a pixel-wise manifold assumption over both the labeled and unlabeled images, in order to enforce a smoothness constraint: *Image pixels that are visually and spatially close to each other are coherent in the semantic space*. We thus propose to learn a manifold [50], [3] for representing all pixels so as to preserve their relative spatial relationships. A classical method, for example, is the Laplacian Embedding [2]. Similar ideas have also been exploited by traditional semi-supervised methods and most recently are integrated with deep leaning representations [44]. In this work, we generalize this methodology to learn a FCN for hierarchal object parsing.
- **Exclusive Constraints**. We introduce a set of exclusive constraints to regularize the FCN network: *image pixels with significant scores from a class (e.g. head) will not be scored significantly for another exclusive class (e.g., torso)*. In multi-label settings, there are multiple exclusive label lists and the labels in each list should be exclusively assigned to an image pixel.

We integrate the above two types of constraints to define a unified multi-label loss function. This function is differentiable with respect to network parameters and hence can be optimized by standard stochastic gradient methods [27]. Our approach can take advantage of both labeled and unlabeled images, providing a simple yet effective way to formulating hierarchical object parsing in semi-supervised settings.

We evaluate the proposed method on two public image datasets and compare it to the alternative object parsing methods. Experiments with comparisons showed that our method can closely match the performance of fully-supervised systems while using only 50% (or less) labeled images. Empirical analysis also validated the effectiveness of the proposed multi-label loss function and regularization terms. Note that we pre-train the proposed method on generic images with classification labels (e.g., ImageNet), without accessing to pixel-wise image labels.

The three **Contributions** of this paper include (i) an effective Multi-label FCN model for hierarchical object parsing that can be trained over both labeled and unlabeled images; (ii) a set of regularization terms, including manifold constraints and exclusive constraints, which are applicable to other image tasks; and (iii) a weakly supervised image parsing system that can achieve state-of-the-art performance while using a small number of fully annotated images.

II. Relationships to Previous Works

The proposed research is closely related to four research streams in computer vision and machine learning.

Object Part Detection has been extensively studied in computer vision literature. The successful deformable part-based model (DPM) [13] [37] and poselet model [5] can effectively represent geometric relationships between object parts in 2D and 3D, but are restricted to their shallow representations while dealing with object instances with large variances. Chen et al. [8] introduced rich contextual part relationships to boost system robustness. Girshick et al. [14] re-formulated the DPM model using convolution neural networks to favor end-to-end learning of deep features. Song et al. [37] proposed to discriminatively train a hierarchical graphical model to allow fined-grained object detection. Wang et al. [43] proposed to jointly segment objects and object parts through learning a two-stream FCN in order to exploit the compositional relationships between part labels. While achieving impressive results, these algorithms did not explicitly formulate cooperative relationships between object parts. For example, an image pixel classified as head should be recognized as 'upper-body as well, not vice versa; or that a pixel should not be simultaneously assigned to upper-body and lower-body. In this work, we will introduce an multi-label loss function to explicitly formulate such coherence and exclusive constraints and use them to guide the learning of deep features.

Semi-supervised methods [50] can be used to train machine learning models using a small number of labeled data. It has made use of embedding techniques [29], which aim to solve a lower-dimensional data representation while preserving pairwise distances in the original feature space. Most embedding algorithms utilize the structure assumption: points within the same structure (or a manifold) are likely to have the same label, and use unlabeled data to discover this structure. Successful approaches include cluster kernels [23], label propagation [30], LapSVM [48], MDSCAL [22], or ISOMAP [38]. Weston et al. [44] employed these embedding methods as regularization terms to learn deep multi-layer neural networks and achieved promising results. Similarly, this work presents a multi-label convolutional network to learn deep features for hierarchical object parsing. Our method explicitly enforces both manifold assumption and coherence/exclusiveness constraints between labels, and showed promising results in reducing the

necessary amount of labeled data to achieve the same level of performance.

Fully Convolution Networks (FCNs) [27] and its variants [7] [31] have been widely used to predict pixel-wise labels and have shown compelling quality and efficiency on multiple datasets [46] [26]. A FCN takes an image as the input, and performs sliding-window-based classification at each pixel in a local receptive field. The network can be trained end-to-end given the pixel-wise semantic region labels. Pinheiro et al. [35] utilizes a FCN to predict object segmentation masks given an input image. To detect object instances, Dai et al. [9] presented an instance-sensitive FCN to generate category-wise instance score maps. All the above models are trained with full supervisions to predict a single label for each pixel but are not suitable for hierarchical object parsing. In this work, we introduce a multi-label FCN in the semi-supervised setting, which is a novel technique in the catalog of hierarchical parsing methods.

Weakly Supervised Image Segmentation Methods in the literature include two major categories. The first category uses image-level labels and automatically reasons segment-label correspondences during training [28], [40]. Popular techniques include the Expectation-Maximization algorithm [11] [31], Probabilistic Generative Models [41], Multi-Instance Methods [32] [34] and Latent Support Vector Machine [24] etc. The second category [45] [49] [16] [6], [25] uses bounding boxes around objects (e.g. humans, animals), instead of pixel-wise image labels, to train image segmentation models. Box inputs are also used as weak supervisions to guide interactive image segmentation [36]. Kumar et al. [24], Xu et al. [47], and Papandreou et al. [31] exploited both image labels and bounding boxes as weak supervisions. In this work, we focus on the semi-supervised image segmentation problem for which a large number of training images are completely unlabeled and aim to learn to hierarchically segment objects in the multi-label setting. We propose to augment the popular FCN model with both traditional semi-supervised regularizations and multi-label exclusive constraints in order to guide the training of deep features.

III. Semi-supervised Multi-label FCN for Hierarchal Object Parsing

In this section we present a learning based hierarchical object parsing algorithm built on top of the expressive Fully Convolution Networks (FCNs) [27]. We focus on methods for training the FCN parameters from both annotated and unannotated images.

Notations We denote by C the total number of object parts and sub parts. These parts form a hierarchy as graphically shown in Figure 2. We denote by $\mathbf{D} = \mathbf{D}^l \cup \mathbf{D}^u$ the set of training images, including labeled images $\mathbf{D}^l$ and unlabeled images $\mathbf{D}^u$. Let $x \in \mathbf{D}$ denote a training image. For a labeled image $x \in \mathbf{D}^l$, we denote by y the segmentation map. Let x_i denote the image pixel at location i, and $y_i \in \{1, \ldots, C\}$ denote its pixel-wise semantic label. We assume that the number of unlabeled samples in $\mathbf{D}^u$ are much larger than that of the labeled samples in $\mathbf{D}^l$. We denote the outputs of the convolutional network by $f(\cdot)$, which can be considered as a scoring function of the image x.

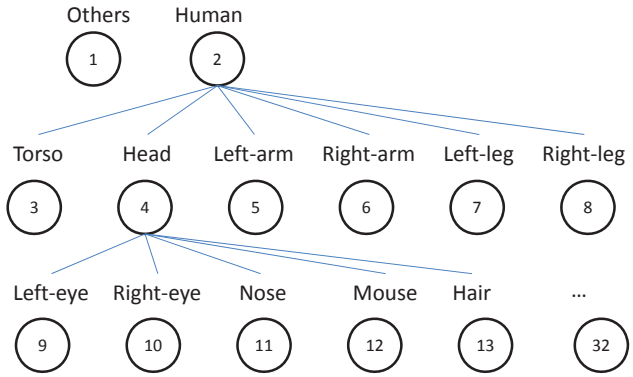

Fig. 2. Hierarchy of human parts used in this paper. There are a total of 31 human parts forming a tree-like structure.

The rest of this section is organized as follows: In Subsection III-A, we formulate the problem of hierarchical object parsing as a multi-label classification problem. In Subsection III-B, we introduce how to extend the proposed formula to the semi-supervised setting. In Subsection III-C, we present a set of regularizations to the proposed semi-supervised model. In Subsection III-D, we specify the unified formula used in this work. In Subsection III-E, we elaborate on the implementation of the proposed algorithm.

A. Multi-label FCN with Unidirectionally Coherent Constraints

In our approach for hierarchical object parsing, we adopted the fully convolutional network (FCN) proposed in [27] as the basic network architecture and investigate ways to specify effective loss functions in the multi-label setting. Our method considers multiple part labels which form a tree-like structure, as shown in Figure 2, and enforces the following coherence constrains: *for any image pixel x_i, the prediction score for label A should be at least equal to the score for any offspring labels of A*. These pair-wise coherence constraints are unidirectional and should be satisfied during the learning of deep features.

We augment the multi-label softmax loss [17], which can be used for single-label predictions as well, with extra regularizations in order to enforce the proposed unidirectional coherence constrains. Let $f_i(k)$ denote the k-th output layer of the FCN Network, which is the activation value for an image pixel x_i and class k, and $\hat{f}_i(k)$ denote corresponding probability, obtained as:

$$\hat{f}_i(k) = \frac{\exp[f_i(k)]}{\sum_{l=1}^{C} \exp[f_i(l)]} \tag{1}$$

Let $\hat{f}_i = [\hat{f}_i(k)], k = 1, 2, \ldots$ assembles the probability of x_i belonging to every label. Let y_i denote a C-dimensional label vector, whose k-th component is 1 if x_i belongs to the class k; 0, otherwise. We normalize y_i so that its sum is unit 1, and use

it as the ground-truth probability. Thus, given a set of labeled images, we aim to learn a FCN so as to minimize the Kullback-Leibler (KL) divergence from the prediction probability $\hat{f}_i$ to ground-truth probability y_i. Such a loss function is also regularized by the between-label coherence constraints. Let $\mathcal{P}(k)$ denote the set of offspring labels of the label k. We define the loss function over labeled images as follows,

$$\mathcal{J}(x,y) = \sum_{i=1}^{n} KL(\hat{f}_i \| y_i) \quad (2)$$

$$s.t., \forall l \in \mathcal{P}(k), \hat{f}_i(k) > \hat{f}_i(l) \quad (3)$$

$$k \in [1, C] \quad (4)$$

where $KL(\hat{f}_i \| y_i) = \sum_k \hat{f}_i(k) \log \frac{\hat{f}_i(k)}{y_i(k)}$. Eq. (2) is a constrained logarithm function and can be optimized using the standard gradient method [21]. This supervised method requires a large amount of labeled data, which is cost-intensive to prepare. In the next subsection, we will introduce a semi-supervised variant to take advantages of large-scale unlabeled images.

B. Manifold Regularization

We adopt the Eq. (2) to the semi-supervised setting in order to learn a multi-label classifier capable of generalizing to unseen testing images. Being similar to most semi-supervised methods [50], we assume that the number of labeled samples is much smaller than that of unlabeled samples, and that pixel-wise features of the same image are drawn from one or multiple manifold subspaces. We employ a Laplacian graph [48] to impose the above manifold regularizations. Let W_{ij} denote the similarity between the image pixels x_i and x_j. We define the Laplacian matrix by $L = W - D$, where $D_{ii} = \sum_j W_{ij}$ is diagonal. Let f_i denote the predicated label vector for the image pixel i and $F = [f_i]$ the predicated label matrix. Thus, we introduce the following regularization term,

$$\mathcal{U}(x) = \sum_i \sum_j W_{ij} \|f_i - f_j\|^2 \quad (5)$$

$$= tr(F^T L F) \quad (6)$$

$$s.t. \quad F^T D F = \mathbf{1}, F^T D \mathbf{1} = 0 \quad (7)$$

where $tr(\cdot)$ represents the trace of a matrix, $\mathbf{1}$ indicates a full-one matrix and the two constraints are used to avoid trivial solutions [50].

C. Integrating Mutually Exclusive Constraints

In the proposed multi-label setting, an image pixel might be assigned to multiple coherent labels, e.g., head and nose. In the meantime, for example, the labels of head and torso should be exclusively assigned to an image pixel. Such exclusive constraints are mutually effective for part labels, and should be satisfied during the training of deep features. Therefore, we regularize the proposed FCN model with the following constraint: *if an image pixel x_i receives a relatively large score for a part label k, it is less likely for x_i to receive significant scores for and only for the part labels that are exclusive from the label k*. Figure 2 graphically shows the decomposition relationships between part labels. The exclusive labels of a part (e.g., arm) include the parts of the same level in the hierarchy (e.g., head) and their offspring parts (e.g., eye, nose).

To enforce the above exclusive constraints, we impose additional regularizations terms over network activities $f(\cdot)$. Let $C(k)$ denote the exclusive labels for the label k. For each label k and image pixel x_i, we utilize an unified function to accumulate the output activities for the labels in $C(k)$, denoted as $\hat{p}(x_i; k)$. We normalize $\hat{p}(x_i; k)$ so that its sum is unit 1. Being similar to [17], we specify a loss regularization over both labeled and unlabeled images, which aims to minimizes the entropy of the distribution $\hat{p}(x_i; k)$:

$$\Omega(x) = -\sum_{k=1}^{C} \sum_{l \in C(k)} \hat{p}_l(x_i; k) \log \hat{p}_l(x_i; k) \quad (8)$$

Minimizing the above entropy will encourage sparsity over the distribution $\hat{p}(x_i; k)$ and thus directly encodes the proposed exclusive constraints.

D. Unified formula: Semi-supervised Multi-label FCN

We define a unified loss function to integrate the objectives in Eqs. (2), (5), and (8):

$$\mathcal{L}(\mathbf{D}) = \sum_{x \in \mathbf{D}^l} \lambda_l \mathcal{J}(x,y) + \sum_{x \in \mathbf{D}} [\lambda_u \mathcal{U}(x) + \lambda_e \Omega(x)] \quad (9)$$

where λ_l, λ_u and λ_e are constants and their sum is 1. Among these terms, $\mathcal{J}(x)$ is the multi-label supervised loss, $\mathcal{U}$ is the manifold regularization tem and $\Omega(x)$ is the regularization tem with exclusive constraints. These objectives are defined over both labeled and unlabeled samples.

E. Network Architecture

Figure 3 graphically illustrates the network architecture of the proposed semi-supervised multi-label fully convolutional network. We use the VGG-16 network architecture, and employ the atrous algorithm [7] to generate dense pixel-wise predictions. These convolutional layers are applied on the input image to get a pixel-wise score map for each part label. We pre-train the network on ImageNet with the cross-entropy loss function [21], and fine-tune the network weights following the procedure of Long et al. [27]. In particular, we replace the 1000-way ImageNet classifier in the last layer of VGG-16 with a 32-way one, corresponding to the 31 human part labels plus background label. On top of this pre-trained network, we replace the last fully connected layers with two convolution layers [9]: one layer uses 1×1 kernels and the other layer uses 3×3 kernels to generate pixel-wise predictions. Like Long et al. [27], we upsample and concatenate the intermediate predictions to get pixel-wise scores and use them calculate losses over training images.

The proposed loss function Eq. (9) is smooth and differentiable, and can be effectively optimized using the standard stochastic gradient descent algorithm [21]. In particular, we run forward propagation on the input image, generating pixel-wise score maps. Each image pixel is associated with a score for each of the 32 labels. It takes a total of 0.18 seconds to evaluate an image on a K40 GPU. With pixel-wise prediction

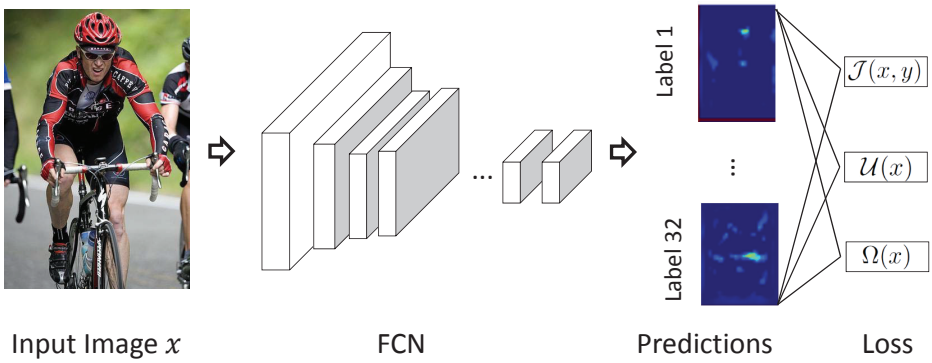

Fig. 3. Network architecture of the proposed semi-supervised multi-label fully convolutional network. The loss functions are defined over both labeled and unlabeled data, and include three parts. $\mathcal{J}$: multi-label softmax loss with unidirectional coherence constraints; $\mathcal{U}$: loss over Laplacian graph; Ω: loss over mutually exclusive constraints.

maps, we use the fully connected Conditional Random Field (CRF) model [7] to obtain the final label assignment. This post-processing step is known to be effective for smoothing regions and refining region boundaries. It is noted that learning potential functions for the CRF models simultaneously can bring extra improvements in performances [7].

IV. Experiments

In this section, we test and evaluate the proposed semi-supervised method for hierarchical object parsing using public image benchmarks and compare it to the other popular methods.

A. Evaluation Protocols

Datasets We use two public datasets for object parsing. The first one is the **UCLA Human Part dataset**, which is a subset of the UCLA PASCAL Part Challenge [46]. The dataset includes 1716 training images and 1817 testing images. It provides pixel-wise part annotations of 31 human parts, as shown in Figure 2. This challenge requires multi-level part recognitions which are more challenging than the alternative benchmarks [21] [26]. The second dataset is the **PASCAL Quadrupseds dataset** [43], which includes images of five animals, including cat, dog, sheep, cow and horse. There are 3120 training images and 294 testing images, annotated with 4 part labels (including head, body, leg and tail). Note that the second dataset is provided with two-level of part annotations: the whole object (level-1) and part labels (level-2), and the proposed multi-label framework is still applicable. These two image datasets include a variety of natural object images, being used to test the generalization capability of the proposed hierarchical object parsing algorithm.

Image Augmentation Using a sufficient amount of representative training images is crucial to the success of deep learning models. In this work, we resize each training image so that its longer dimension is 500 pixels, and slide a subwindow of 300 by 300 pixels with a step size of 20 pixels. For each subwindow, we perform 4 additional croppings through randomly selecting one of the following ways: (i) flipping, with a probability 0.1; (ii) changing color intensity by a random scale in [0.7, 1.3], with a probability 0.4; (iii) rotating a random degree between [-5, 5], with a probability 0.5. We cropped 30-70 samples for each image. Similar cropping protocol has been used in previous works, e.g., [43].

Implementation To measure the pair-wise distance between image pixels, we extract the Histogram of Oriented Gradient (HOGs) [10] from local regions centered at individual pixels, and calculate their pair-wise Euclidean distance W_{ij}. In the loss function Eq 9, we set λ_l, λ_u, and λ_e to be 0.6, 0.3, and 0.1, respectively. We train the semi-supervised multi-label FCN using stochastic gradient descent methods with mini-batches. Each mini-batch contains 30 images. The initial learning rate is 0.001 and is decreased by a factor of 0.1 after every 2000 iterations. We set the momentum to be 0.9 and the weight decay to be 0.0005. The initialization model is a modified VGG-16 network pre-trained on ImageNet. Fine-tuning our network on the first UCLA Human Part dataset takes about 30 hours on a NVIDIA Telsa K40 GPU. The average inference time for one image is about 0.3 seconds. While applying the proposed method over each of the three datasets, we use the $10,582$ images from PASCAL VOC 2012 as unlabeled images. Being similarly to [7], we decouple the DCNN and Dense CRF training stages and learn the CRF parameters by cross validation to maximize IOU (intersection-over-union) segmentation accuracy in a heldout set of 100 fully-annotated images. We use 10 mean-field iterations for Dense CRF inference [20].

Baselines We compare our algorithm with three state-of-the-art methods for **part segmentation**, including two popular supervised methods: the Deep Hypercolumn (HC) method [19], and the Joint Object and Part Segmentation method by Wang et al. [43]. Both methods require fully annotated training images. We also compare to the weakly supervised method by Papandreou et al. [31] which can automatically infer pixel-wise segmentation maps using an EM method. We implemented and trained their models following the suggested configurations/procedures.

Evaluation Metrics We evaluate the results of various object parsing methods using IOU, i.e. intersection-over-union between pixel-wise predictions and ground-truth labels. The

proposed method might generate multiple label predictions for every pixel. In the evaluation, we consider each part/subpart as a separate class, and compute IOU for each class. We calculate the mean IOU across images and average over all labels.

B. Results on the UCLA Human Part Dataset [46]

We apply the proposed semi-supervised method over the UCLA Human Part dataset and evaluate it in both inductive and transductive settings. The former studies how well the learned model works on unseen examples, while the latter studies how the learning procedure discovers labels for the unlabeled training samples [50].

Table I reports the quantitative comparisons of all methods in the inductive setting. We learn the proposed model from both labeled and unlabeled data and test the learned model over unseen testing samples. Among the 31 part labels, we did not include the results for the six subparts of head (e.g., ear, eye, mouse, brow, nose, hair) since their instances are very rare. We evaluated three variants of the proposed method: Our-I, Our-II, Our-III, which used 30%, 50%, and 100% labeled training samples, respectively. The three baseline methods used 100% labeled training samples for fair comparisons since the three implementations of our method access extra unlabeled images. The semi-supervised method [31] used all the unlabeled images. We implement another variant of the proposed method, denoted as Our-IV, which does not utilize exclusive label constraints. We set $\lambda_u = 0.4$ and $\lambda_e = 0$ for Our-IV. We use these variants to analyze the effects of individual components of the proposed method.

From the comparisons of various labeling algorithms, we can draw the following observations. *First*, the proposed method can achieve equivalent performance to the three fully supervised object parsing methods while only using 50% or less labeled images. With only 30% labeled training images, the proposed Our-I (IOU: 51.4%) can still outperform the methods [19] (IOU: 48.3%) and [43](IOU: 50.2%), which is an encouraging result considering it is cost-intensive to collect hierarchical part annotations. Our semi-supervised method, however, still needs a descent amount of supervisions to be properly trained. Note that our method is different from one-short learning algorithms [12] [42] that work on one or a few labeled training samples. *Second*, with additional use of unlabeled images the semi-supervised methods Our-III and [31] can achieve equivalent performance as the fully supervised methods. This observation is consistent with the previous works [31] and demonstrates the great potential of semi-supervised methods in conjunction with advanced deep learning techniques. *Third*, the comparisons between Our-IV and Our-III demonstrated the advantages of the proposed exclusive constraints. In particular, our method obtained about 3 percentages improvements while additionally employing such constraints.

Figure 4 visualizes exemplar results of transductive inference using the proposed method Our-I. For each unlabeled training image, we run forward propagation over the learned network to get label-wise prediction maps and employ the densely connected CRF method [7] to obtain final labels. For each image in the first-column, we visualize its label map using using three figures for clarity. Column 1: human and background; Column 2: torso, head, neck, arm, and leg; Column 3: other part labels. Note that we change the color codes used in different columns to highlight the semantic regions obtained. These images include many challenges to existing state of object parsing, including occlusions (row 1), complex interactions (row 2), lighting changes (row 3) and scale change (row 4). With the proposed constraints, our method achieved promising results considering that only a small number of labeled images is used for training.

C. Results on the Quadrupeds dataset [43]

We further test and evaluate the proposed method over the Quadrupeds dataset. We used the same baseline methods as the previous experiment. Table III reports the quantitative comparisons between all algorithms using IOU metrics. We first calculate IOU for each label, and then average across part labels to get the category-wise IOU. For every method, we also average category-wise IOU over all object categories for comparisons. The comparisons between various methods clearly demonstrate the advantages of the proposed method. Notably, with 30% labeled training images, the proposed method Our-I can achieve much better performance (49.6%) than two state-of-the-art methods [19] (38.8%) and [43] (44.3%). It is also comparable to the semi-supervised method [31] (51.7%), which uses all the labeled training images. Moreover, we can observe that Our-IV achieved a decent accuracy (52.0%) while only using multi-label loss and Laplacian regularization, and obtained a much better accuracy (56.6%) while additionally using the proposed exclusive constraints.

Fig. 5 visualizes the results of various human part parsing methods, including [19], [43] and Our-I. The three exemplar images include cat, horse and cow, respectively. We show the ground-truth label map in the last column for comparisons. For the image of cat (row-1), [19] is less accurate than the other two methods since the labeled region of legs include many background pixels. For the other two images, only the proposed method can identify the part of tail. Our method aims to directly model the coherence and exclusive relationships in the label space, and demonstrates much stronger generalization capability than the alternatives.

V. Conclusions

This paper presented a multi-label fully convolutional network (FCN) that can be effectively trained on semi-supervised images for hierarchical object parsing. Our model is capable of explicitly imposing the various constraints between image labels and taking advantages of unlabeled images. In particular, we introduced three types of constraints: (i) Pair-wise coherences between part labels and their offspring labels, (ii) Pixel-wise manifold regularization, and (iii) Exclusive constraints between object parts labels. We formulated these objectives in a unified loss function and use it to learn deep features in the semi-supervised setting. The proposed model can be end-to-end trained using the standard stochastic gradient algorithm.

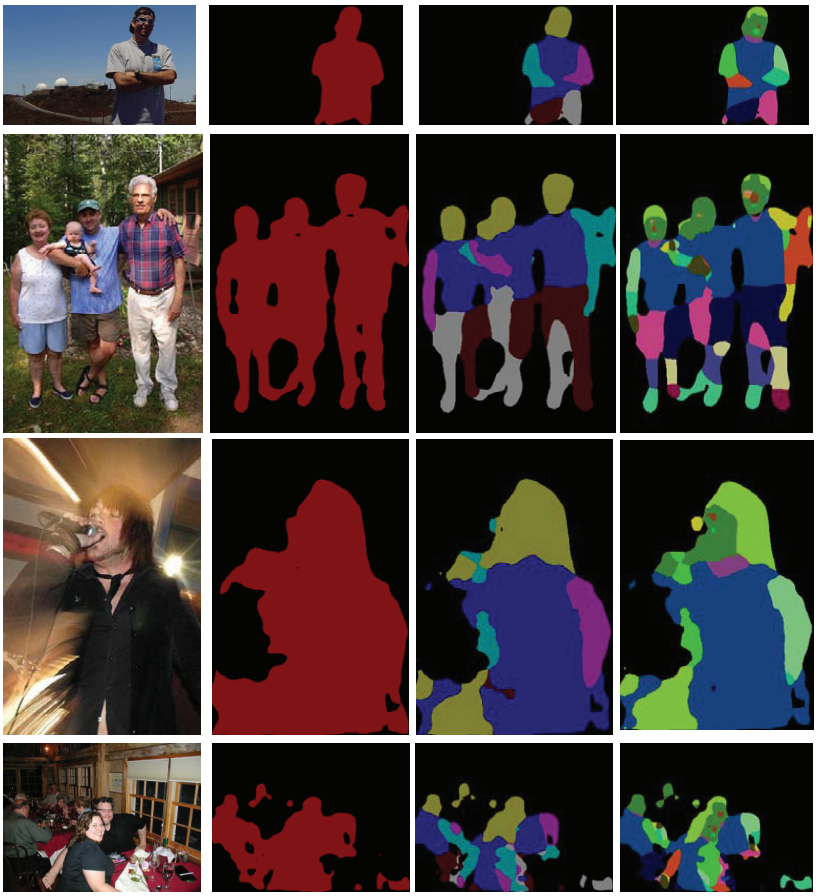

Fig. 4. Exemplar results of transductive learning. Column-1 shows the input images. Columns 2-4 show the predicted maps for binary labels (human and others), five part labels (torso, head, neck, arm, leg), and the other subpart labels, respectively.

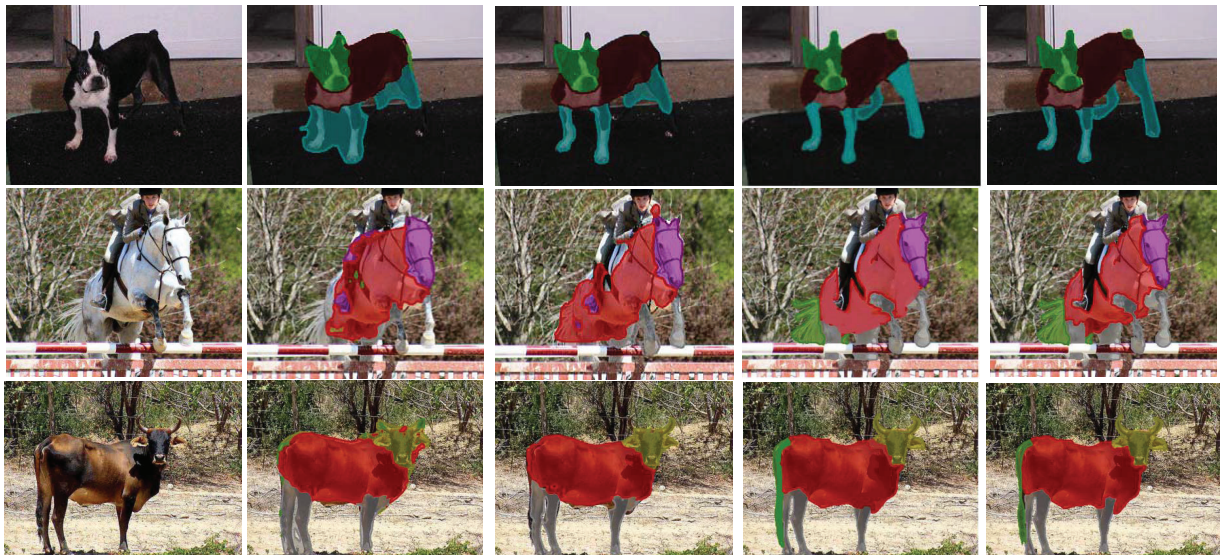

Fig. 5. Exemplar results of part parsing. Column 1: Input images; Column 2: results by [19]; Column 3: Results by [43] ; Column 4: Results by the proposed method (Our-I); Column 5: groundtruth label map. Color codes are randomly generated to highlight the segmented regions.

	human	torso	head	neck	arm	leg	u-arm	l-arm	hand	u-leg	l-leg	foot	Avg.
[19]	77.6	78.2	76.2	38.5	41.7	35.4	37.6	53.1	48.9	38.1	25.4	29.3	48.3
[43]	79.4	81.2	75.8	42.8	42.5	38.8	39.3	54.3	50.4	38.7	27.2	31.9	50.2
[31]	81.9	82.4	81.2	40.3	43.2	42.5	48.2	61.2	51.8	42.7	29.4	31.4	53.0
Our-I	78.3	79.3	80.4	41.6	39.1	41.2	46.1	59.1	50.5	36.5	32.4	32.7	51.4
Our-II	83.5	81.4	82.5	43.3	41.3	43.7	50.2	63.7	56.3	45.1	33.5	34.5	54.9
Our-III	**86.1**	**84.5**	**87.9**	**45.2**	**47.8**	**52.1**	**55.4**	**65.2**	**58.4**	**48.5**	**36.5**	**42.1**	**59.1**
Our-IV	83.1	82.3	84.7	44.1	42.3	48.5	51.4	62.8	54.3	47.6	34.1	38.7	56.0

TABLE I

PART PARSING RESULTS (IOU) ON THE UCLA HUMAN PART CHALLENGE [46]. THE PROPOSED SEMI-SUPERVISED METHOD OUR-I, OUR-II, AND OUR-III USED 30%, 50%, AND 100% ANNOTATED TRAINING SAMPLES, RESPECTIVELY. THE METHOD OUR-IV USES 100% ANNOTATED IMAGES BUT DOES NOT EMPLOYS THE EXCLUSIVE CONSTRAINTS BETWEEN LABELS. THE OTHER BASELINE METHODS USE ALL THE TRAINING SAMPLES.

	Dog	Cat	Cow	Horse	Sheep	Avg.
[19]	42.1	44.0	35.5	38.6	33.8	38.8
[43]	45.6	47.8	42.7	49.6	35.7	44.3
[31]	54.2	53.4	46.8	55.7	44.1	50.8
Our-I	50.3	54.2	45.9	53.1	44.3	49.6
Our-II	52.4	55.6	46.7	55.8	48.1	51.7
Our-III	**57.9**	**63.2**	**52.4**	**58.3**	**55.1**	**57.4**
Our-IV	52.4	57.4	46.1	56.3	47.8	52.0

TABLE II

PART PARSING RESULTS (IOU) ON THE QUADRUPEDS DATASET. THE PROPOSED SEMI-SUPERVISED METHOD OUR-I, OUR-II, AND OUR-III USED 30%, 50%, AND 100% ANNOTATED TRAINING IMAGES, RESPECTIVELY. THE OTHER BASELINE METHODS USED ALL THE LABELED TRAINING IMAGES. THE METHOD OUR-IV USES 100% ANNOTATED IMAGES BUT DOES NOT EMPLOYS THE EXCLUSIVE CONSTRAINTS BETWEEN LABELS.

	Dog	Cat	Cow	Horse	Sheep	Avg.
[19]	42.1	44.0	35.5	38.6	33.8	38.8
[43]	45.6	47.8	42.7	49.6	35.7	44.3
[31]	54.2	53.4	46.8	55.7	44.1	50.8
Our-I	50.3	54.2	45.9	53.1	44.3	49.6
Our-II	52.4	55.6	46.7	55.8	48.1	51.7
Our-III	**57.9**	**63.2**	**52.4**	**58.3**	**55.1**	**57.4**
Our-IV	52.4	57.4	46.1	56.3	47.8	52.0

TABLE III

RESULTS (IOU) ON THE QUADRUPEDS DATASET. THE PROPOSED SEMI-SUPERVISED METHOD OUR-I, OUR-II, AND OUR-III USED 30%, 50%, AND 100% ANNOTATED TRAINING IMAGES, RESPECTIVELY. THE BASELINE METHODS USED ALL THE LABELED TRAINING IMAGES. THE METHOD OUR-IV USES 100% ANNOTATED IMAGES BUT DOES NOT EMPLOYS THE EXCLUSIVE CONSTRAINTS BETWEEN LABELS.

Experiments with comparisons on public image datasets showed that our method can achieve state-of-the results for segmenting object parts of varying semantic levels in images. We empirically showed that: (i) Additional use of a large amount of unlabeled images brought significant improvements in multi-level part segmentation; (ii) Our method achieved comparable performance while using 50% or less labeled samples than the alternatives; (iii) The proposed coherence constrains and exclusive constraints resulted in improved performance, respectively, and the integration of these two constraints achieve state-of-the-art performance for semi-supervised hierarchical object parsing.

REFERENCES

[1] Mohamed R Amer, Dan Xie, Mingtian Zhao, Sinisa Todorovic, and Song-Chun Zhu. Cost-sensitive top-down/bottom-up inference for multi-scale activity recognition. In *European Conference on Computer Vision*, pages 187–200. Springer, 2012.

[2] Mikhail Belkin and Partha Niyogi. Laplacian eigenmaps and spectral techniques for embedding and clustering. In *NIPS*, volume 14, pages 585–591, 2001.

[3] Mikhail Belkin, Partha Niyogi, and Vikas Sindhwani. Manifold regularization: A geometric framework for learning from labeled and unlabeled examples. *Journal of machine learning research*, 7(Nov):2399–2434, 2006.

[4] Yoshua Bengio, Aaron Courville, and Pascal Vincent. Representation learning: A review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence*, 35(8):1798–1828, 2013.

[5] Lubomir Bourdev, Subhransu Maji, Thomas Brox, and Jitendra Malik. Detecting people using mutually consistent poselet activations. In *European conference on computer vision*, pages 168–181. Springer, 2010.

[6] Liang-Chieh Chen, Sanja Fidler, Alan L Yuille, and Raquel Urtasun. Beat the mturkers: Automatic image labeling from weak 3d supervision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3198–3205, 2014.

[7] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Semantic image segmentation with deep convolutional nets and fully connected crfs. *arXiv preprint arXiv:1412.7062*, 2014.

[8] Xianjie Chen, Roozbeh Mottaghi, Xiaobai Liu, Sanja Fidler, Raquel Urtasun, and Alan Yuille. Detect what you can: Detecting and representing objects using holistic models and body parts. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1971–1978, 2014.

[9] Jifeng Dai, Kaiming He, Yi Li, Shaoqing Ren, and Jian Sun. Instance-sensitive fully convolutional networks. In *European Conference on Computer Vision*, pages 534–549. Springer, 2016.

[10] Navneet Dalal and Bill Triggs. Histograms of oriented gradients for human detection. In *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, volume 1, pages 886–893. IEEE, 2005.

[11] Pinar Duygulu, Kobus Barnard, Joao FG de Freitas, and David A Forsyth. Object recognition as machine translation: Learning a lexicon for a fixed image vocabulary. In *European conference on computer vision*, pages 97–112. Springer, 2002.

[12] Li Fei-Fei, Rob Fergus, and Pietro Perona. One-shot learning of object categories. *IEEE transactions on pattern analysis and machine intelligence*, 28(4):594–611, 2006.

[13] Pedro Felzenszwalb, David McAllester, and Deva Ramanan. A discriminatively trained, multiscale, deformable part model. In *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, pages 1–8. IEEE, 2008.

[14] Ross Girshick, Forrest Iandola, Trevor Darrell, and Jitendra Malik. Deformable part models are convolutional neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 437–446, 2015.

[15] Ross Girshick, Jamie Shotton, Pushmeet Kohli, Antonio Criminisi, and Andrew Fitzgibbon. Efficient regression of general-activity human poses from depth images. In *Computer Vision (ICCV), 2011 IEEE International Conference on*, pages 415–422. IEEE, 2011.

[16] Matthieu Guillaumin, Daniel Küttel, and Vittorio Ferrari. Imagenet auto-annotation with segmentation propagation. *International Journal of Computer Vision*, 110(3):328–348, 2014.

[17] Matthieu Guillaumin, Thomas Mensink, Jakob Verbeek, and Cordelia Schmid. Tagprop: Discriminative metric learning in nearest neighbor models for image auto-annotation. In *Computer Vision, 2009 IEEE 12th International Conference on*, pages 309–316. IEEE, 2009.

[18] Feng Han and Song-Chun Zhu. Bottom-up/top-down image parsing by attribute graph grammar. In *Computer Vision, 2005. ICCV 2005. Tenth IEEE International Conference on*, volume 2, pages 1778–1785. IEEE, 2005.

[19] Bharath Hariharan, Pablo Arbeláez, Ross Girshick, and Jitendra Malik. Hypercolumns for object segmentation and fine-grained localization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 447–456, 2015.

[20] Vladlen Koltun. Efficient inference in fully connected crfs with gaussian edge potentials. *Adv. Neural Inf. Process. Syst*, 2(3):4, 2011.

[21] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.

[22] Joseph B Kruskal. Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. *Psychometrika*, 29(1):1–27, 1964.

[23] Brian Kulis, Sugato Basu, Inderjit Dhillon, and Raymond Mooney. Semi-supervised graph clustering: a kernel approach. *Machine learning*, 74(1):1–22, 2009.

[24] M Pawan Kumar, Haithem Turki, Dan Preston, and Daphne Koller. Learning specific-class segmentation from diverse data. In *Computer Vision (ICCV), 2011 IEEE International Conference on*, pages 1800–1807. IEEE, 2011.

[25] Victor Lempitsky, Pushmeet Kohli, Carsten Rother, and Toby Sharp. Image segmentation with a bounding box prior. In *Computer Vision, 2009 IEEE 12th International Conference on*, pages 277–284. IEEE, 2009.

[26] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European Conference on Computer Vision*, pages 740–755. Springer, 2014.

[27] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3431–3440, 2015.

[28] Wei-Lwun Lu, Jo-Anne Ting, James J Little, and Kevin P Murphy. Learning to track and identify players from broadcast sports videos. *IEEE transactions on pattern analysis and machine intelligence*, 35(7):1704–1716, 2013.

[29] Feiping Nie, Dong Xu, Ivor Wai-Hung Tsang, and Changshui Zhang. Flexible manifold embedding: A framework for semi-supervised and unsupervised dimension reduction. *IEEE Transactions on Image Processing*, 19(7):1921–1932, 2010.

[30] Zheng-Yu Niu, Dong-Hong Ji, and Chew Lim Tan. Word sense disambiguation using label propagation based semi-supervised learning. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, pages 395–402. Association for Computational Linguistics, 2005.

[31] George Papandreou, Liang-Chieh Chen, Kevin P Murphy, and Alan L Yuille. Weakly-and semi-supervised learning of a deep convolutional network for semantic image segmentation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1742–1750, 2015.

[32] Deepak Pathak, Evan Shelhamer, Jonathan Long, and Trevor Darrell. Fully convolutional multi-class multiple instance learning. *arXiv preprint arXiv:1412.7144*, 2014.

[33] Mingtao Pei, Yunde Jia, and Song-Chun Zhu. Parsing video events with goal inference and intent prediction. In *Computer vision (iccv), 2011 ieee international conference on*, pages 487–494. IEEE, 2011.

[34] Pedro O Pinheiro and Ronan Collobert. 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, 2015.

[35] Pedro O Pinheiro, Ronan Collobert, and Piotr Dollar. Learning to segment object candidates. In *Advances in Neural Information Processing Systems*, pages 1990–1998, 2015.

[36] Carsten Rother, Vladimir Kolmogorov, and Andrew Blake. Grabcut: Interactive foreground extraction using iterated graph cuts. In *ACM transactions on graphics (TOG)*, volume 23, pages 309–314. ACM, 2004.

[37] Xi Song, Tianfu Wu, Yunde Jia, and Song-Chun Zhu. Discriminatively trained and-or tree models for object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3278–3285, 2013.

[38] Joshua B Tenenbaum, Vin De Silva, and John C Langford. A global geometric framework for nonlinear dimensionality reduction. *science*, 290(5500):2319–2323, 2000.

[39] Zhuowen Tu, Xiangrong Chen, Alan L Yuille, and Song-Chun Zhu. Image parsing: Unifying segmentation, detection, and recognition. *International Journal of computer vision*, 63(2):113–140, 2005.

[40] Jakob Verbeek and Bill Triggs. Region classification with markov field aspect models. In *Computer Vision and Pattern Recognition, 2007. CVPR'07. IEEE Conference on*, pages 1–8. IEEE, 2007.

[41] Alexander Vezhnevets, Vittorio Ferrari, and Joachim M Buhmann. Weakly supervised structured output learning for semantic segmentation. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 845–852. IEEE, 2012.

[42] Oriol Vinyals, Charles Blundell, Tim Lillicrap, Daan Wierstra, et al. Matching networks for one shot learning. In *Advances in Neural Information Processing Systems*, pages 3630–3638, 2016.

[43] Peng Wang, Xiaohui Shen, Zhe Lin, Scott Cohen, Brian Price, and Alan L Yuille. Joint object and part segmentation using deep learned potentials. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1573–1581, 2015.

[44] Jason Weston, Frédéric Ratle, Hossein Mobahi, and Ronan Collobert. Deep learning via semi-supervised embedding. In *Neural Networks: Tricks of the Trade*, pages 639–655. Springer, 2012.

[45] Wei Xia, Csaba Domokos, Jian Dong, Loong-Fah Cheong, and Shuicheng Yan. Semantic segmentation without annotating segments. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2176–2183, 2013.

[46] Peng Wang Xiaochen Lian Junhua Mao Alan Yuille Seong-Whan Lee Xiaobai Liu, Nam-Gyu Cho. Pascal semantic part: Dataset and benchmark. 2014.

[47] Jia Xu, Alexander G Schwing, and Raquel Urtasun. Learning to segment under various forms of weak supervision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3781–3790, 2015.

[48] Shuicheng Yan, Dong Xu, Benyu Zhang, Hong-Jiang Zhang, Qiang Yang, and Stephen Lin. Graph embedding and extensions: A general framework for dimensionality reduction. *IEEE transactions on pattern analysis and machine intelligence*, 29(1), 2007.

[49] Jun Zhu, Junhua Mao, and Alan L Yuille. Learning from weakly supervised data by the expectation loss svm (e-svm) algorithm. In *Advances in Neural Information Processing Systems*, pages 1125–1133, 2014.

[50] Xiaojin Zhu. Semi-supervised learning literature survey. 2005.