Learning Semisupervised Multilabel Fully Convolutional Network for Hierarchical Object Parsing

Xiaobai Liu, Qian Xu, Grayson Adkins, Eric Medwedeff, Liang Lin, *Senior Member, IEEE*, and Shuicheng Yan

***Abstract*— This article presents a semisupervised multilabel 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 multilabel softmax loss function over both labeled and unlabeled images and regularizing it with two pairwise 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 receives significant scores from more than one label 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 data sets for hierarchical object parsing and compare it with 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 (FCN), hierarchical models, semisupervised learning.**

Manuscript received July 24, 2017; revised May 2, 2018 and November 19, 2018; accepted April 1, 2019. Date of publication December 23, 2019; date of current version July 7, 2020. The work of X. Liu was supported in part by the National Science Foundation under Grant 1657600, in part by the Office of Naval Research (ONR) under Grant N00014-17-1-2867, and in part by the San Diego State University Presidential Leadership Funds. *(Corresponding author: Xiaobai Liu.)*

X. Liu is with the Department of Computer Science, San Diego State University, San Diego, CA 92150 USA (e-mail: xiaobai.liu@mail.sdsu.edu).

Q. Xu is with Xrelab Inc., San Diego, CA 92128 USA, and also with the Department of Computer Science, San Diego State University, San Diego, CA 92150 USA.

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

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

L. Lin is with the Human Cyber Physical Intelligence Integration Laboratory, Sun Yat-sen University, Guangzhou 510275, China.

S. Yan is with the Department of Computer and Electrical Engineering, National University of Singapore, Singapore 119077.

Color versions of one or more of the figures in this article are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TNNLS.2019.2931183

I. Introduction

THE goal of this article is to develop an effective approach capable of segmenting objects, object parts (e.g., head, torso, and legs), and subparts (e.g., eyes and mouses) in images and generating a hierarchical representation of objects. The outcomes of our approach include a pixelwise 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 [13], [14], [18], [37], [39], 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 multilevel feature representations for semantic image segmentation. Multiple object parsing methods [9], [31], [43] 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 shows the 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 Fig. 1(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 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 multilabel FCN from images. Hierarchical object parsing is essentially a multilabel image segmentation problem, which aims to assign each pixel of the input image to multiple part labels, e.g., eye,

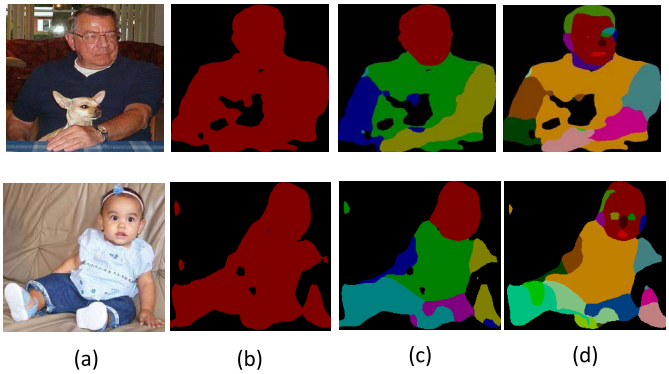

Fig. 1. Hierarchical object parsing. (a) Input images. (b)–(d) Object masks segmented by the proposed method for 2 labels (human and nonhuman), 6 human part labels, and 31 human subpart labels, respectively.

head, and upper body. While the conventional neural network methods [27] can be used for multilabel 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 article, we propose to develop a multilabel 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 semisupervised setting allows our model to generalize to unseen data samples and avoid potential overfitting with the training data. Overfitting 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 overfitting, in this article, we develop two regularization terms for the proposed semisupervised model.

1) *Manifold Regularization:* We introduce a pixelwise 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 [3], [50] 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 semisupervised methods and most recently are integrated with deep leaning representations [44]. In this article, we generalize this methodology to learn an FCN for hierarchal object parsing.
2) *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 multilabel 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-mentioned two types of constraints to define a unified multilabel 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 semisupervised settings.

We evaluate the proposed method on two public image data sets 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 multilabel loss function and regularization terms. Note that we pretrain the proposed method on generic images with classification labels (e.g., ImageNet), without accessing to pixelwise image labels.

The three contributions of this article include: 1) an effective multilabel FCN model for hierarchical object parsing that can be trained over both labeled and unlabeled images; 2) a set of regularization terms, including manifold constraints and exclusive constraints, which are applicable to other image tasks; and 3) 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 2-D and 3-D 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] reformulated 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 article, we will introduce a multilabel loss function to explicitly formulate such coherence and exclusive constraints and use them to guide the learning of deep features.

Semisupervised 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], Lap support vector machine (SVM) [48], multidimensional scaling method (MDSCAL) [22], or isometric feature mapping (ISOMAP) [38]. Weston *et al*. [44] employed these embedding methods as regularization terms to learn deep multilayer neural networks and achieved promising results. Similarly, this article presents a multilabel 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.

FCNs [27] and its variants [7], [31] have been widely used to predict pixelwise labels and have shown compelling quality and efficiency on multiple data sets [26], [46]. An 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 pixelwise semantic region labels. Pinheiro and Collobert [35] utilize an 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 categorywise instance score maps. All the above-mentioned models are trained with full supervisions to predict a single label for each pixel but are not suitable for hierarchical object parsing. In this article, we introduce a multilabel FCN in the semisupervised 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], multiinstance methods [32], [34], and latent SVM [24]. The second category [6], [16], [25], [45], [49] uses bounding boxes around objects (e.g., humans and animals), instead of pixelwise 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 article, we focus on the semisupervised image segmentation problem for which a large number of training images are completely unlabeled and aim to learn to hierarchically segment objects in the multilabel setting. We propose to augment the popular FCN model with both traditional semisupervised regularizations and multilabel exclusive constraints in order to guide the training of deep features.

III. Semisupervised Multilabel FCN for Hierarchal Object Parsing

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

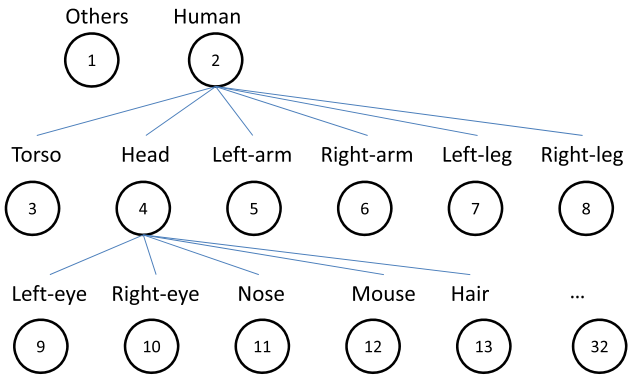

Fig. 2. Hierarchy of human parts used in this article. There are a total of 31 human parts forming a treelike structure.

A. Notations

We denote by C the total number of object parts and subparts. These parts form a hierarchy as graphically shown in Fig. 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 pixelwise semantic label. We assume that the number of unlabeled samples in $\mathbf{D}^u$ is 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.

The rest of this section is organized as follows. In Section III-B, we formulate the problem of hierarchical object parsing as a multilabel classification problem. In Section III-C, we introduce how to extend the proposed formula to the semisupervised setting. In Section III-D, we present a set of regularizations to the proposed semisupervised model. In Section III-E, we specify the unified formula used in this article. In Section III-F, we elaborate on the implementation of the proposed algorithm.

B. Multilabel FCN With Unidirectionally Coherent Constraints

In our approach for hierarchical object parsing, we adopted the FCN proposed in [27] as the basic network architecture and investigate ways to specify effective loss functions in the multilabel setting. Our method considers multiple part labels that form a treelike structure, as shown in Fig. 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 pairwise coherence constraints are unidirectional and should be satisfied during the learning of deep features.

We augment the multilabel 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 kth output layer of the FCN Network, which is the activation value for an image pixel x_i and class k, and $\hat{f}_i(k)$ denotes the corresponding probability,

obtained as

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

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 kth 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:

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

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

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

where $\mathrm{KL}(\hat{f}_i \| y_i) = \sum_k \hat{f}_i(k) \log(\hat{f}_i(k)/y_i(k))$. Equation (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 Section III-C, we will introduce a semisupervised variant to take advantages of large-scale unlabeled images.

C. Manifold Regularization

We adopt (2) to the semisupervised setting in order to learn a multilabel classifier capable of generalizing to unseen testing images. Being similar to most semisupervised methods [50], we assume that the number of labeled samples is much smaller than that of unlabeled samples, and that pixelwise 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 $F = [f_i]$ denote the predicated label matrix. Thus, we introduce the following regularization term:

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

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

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

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

D. Integrating Mutually Exclusive Constraints

In the proposed multilabel 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. Fig. 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 and nose).

To enforce the above exclusive constraints, we impose additional regularizations terms over network activities $f(\cdot)$. Let $\mathcal{C}(k)$ 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 $\mathcal{C}(k)$, 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 $\hat{p}(x_i; k)$

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

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

E. Unified Formula: Semisupervised Multilabel FCN

We define a unified loss function to integrate the objectives in (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 multilabel supervised loss, $\mathcal{U}$ is the manifold regularization term, and $\Omega(x)$ is the regularization term with exclusive constraints. These objectives are defined over both labeled and unlabeled samples.

F. Network Architecture

Fig. 3 graphically illustrates the network architecture of the proposed semisupervised multilabel FCN. We use the VGG-16 network architecture and employ the atrous algorithm [7] to generate dense pixelwise predictions. These convolutional layers are applied on the input image to get a pixelwise score map for each part label. We pretrain 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 pretrained 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 pixelwise predictions. Like Long *et al*. [27], we upsample

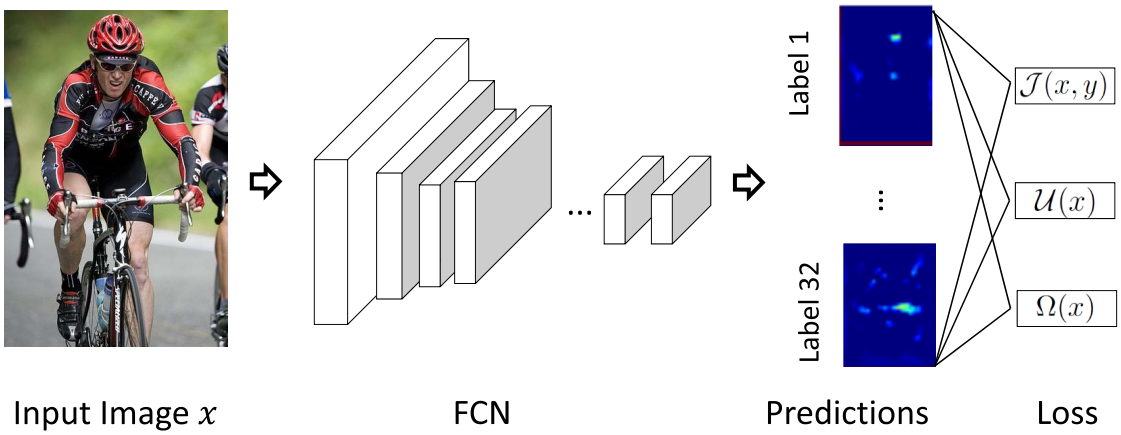

Fig. 3. Network architecture of the proposed semisupervised multilabel FCN. The loss functions are defined over both labeled and unlabeled data and include three parts. $\mathcal{J}$: multilabel softmax loss with unidirectional coherence constraints, $\mathcal{U}$: loss over Laplacian graph, and Ω: loss over mutually exclusive constraints.

and concatenate the intermediate predictions to get pixelwise scores and use them to calculate losses over training images.

The proposed loss function (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 pixelwise score maps. Each image pixel is associated with a score for each of the 32 labels. It takes a total of 0.18 s to evaluate an image on a K40 GPU. With pixelwise prediction maps, we use the fully connected conditional random field (CRF) model [7] to obtain the final label assignment. This postprocessing 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 semisupervised method for hierarchical object parsing using public image benchmarks and compare it to the other popular methods.

A. Evaluation Protocols

1) Data Sets: We use two public data sets for object parsing. The first one is the *UCLA Human Part data set*, which is a subset of the UCLA PASCAL Part Challenge [46]. The data set includes 1716 training images and 1817 testing images. It provides pixelwise part annotations of 31 human parts, as shown in Fig. 2. This challenge requires multilevel part recognitions, which are more challenging than the alternative benchmarks [21], [26]. The second data set is the *PASCAL Quadrupseds data set* [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 data set is provided with two-level of part annotations: the whole object (level-1) and part labels (level-2), and the proposed multilabel framework is still applicable. These two image data sets include a variety of natural object images, being used to test the generalization capability of the proposed hierarchical object parsing algorithm.

2) Image Augmentation: Using a sufficient amount of representative training images is crucial to the success of deep learning models. In this article, 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: 1) flipping, with a probability 0.1; 2) changing color intensity by a random scale in [0.7, 1.3], with a probability 0.4; and 3) 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 (see [43]).

3) Implementation: To measure the pairwise distance between image pixels, we extract the histogram of oriented gradient (HOGs) [10] from local regions centered at individual pixels and calculate their pairwise Euclidean distance W_{ij}. In the loss function 9, we set λ_l, λ_u, and λ_e to be 0.6, 0.3, and 0.1, respectively. We train the semisupervised multilabel FCN using stochastic gradient descent methods with minibatches. Each minibatch 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 pretrained on ImageNet. Fine-tuning our network on the first UCLA Human Part data set takes about 30 h on a NVIDIA Telsa K40 GPU. The average inference time for one image is about 0.3 s. While applying the proposed method over each of the three data sets, we use the 10 582 images from PASCAL VOC 2012 as unlabeled images. Being similar to [7], we decouple the deep convolutional neural network (DCNN) and Dense CRF training stages and learn the CRF parameters by cross validation to maximize intersection-over-union (IOU) segmentation accuracy in a held-out set of 100 fully annotated images. We use 10 mean-field iterations for dense CRF inference [20].

4) Baselines: We compare our algorithm with three state-of-the-art methods for *part segmentation*, including two

TABLE I

PART PARSING RESULTS (IOU) ON THE UCLA HUMAN PART CHALLENGE [46]. THE PROPOSED SEMISUPERVISED 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 EMPLOY THE EXCLUSIVE CONSTRAINTS BETWEEN LABELS. THE OTHER BASELINE METHODS USE ALL THE TRAINING SAMPLES

	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 II

PART PARSING RESULTS (IOU) ON THE QUADRUPEDS DATA SET. THE PROPOSED SEMISUPERVISED 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 EMPLOY 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

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 pixelwise segmentation maps using an electromagnetic (EM) method. We implemented and trained their models following the suggested configurations/procedures.

5) Evaluation Metrics: We evaluate the results of various object parsing methods using IOU, i.e., IOU between pixelwise 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 Data Set [46]

We apply the proposed semisupervised method over the UCLA Human Part data set 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, and hair) since their instances are very rare. We evaluated three variants of the proposed method: Our-I, Our-II, and 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 semisupervised 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.

1) 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 semisupervised 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.
2) With additional use of unlabeled images the semisupervised 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 semisupervised methods in conjunction with advanced deep learning techniques.
3) The comparisons between Our-IV and Our-III demonstrated the advantages of the proposed exclusive constraints. In particular, our method obtained about 3% improvements while additionally employing such constraints.

Fig. 4 visualizes the 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 labelwise prediction maps and employ the densely connected CRF method [7] to obtain final labels.

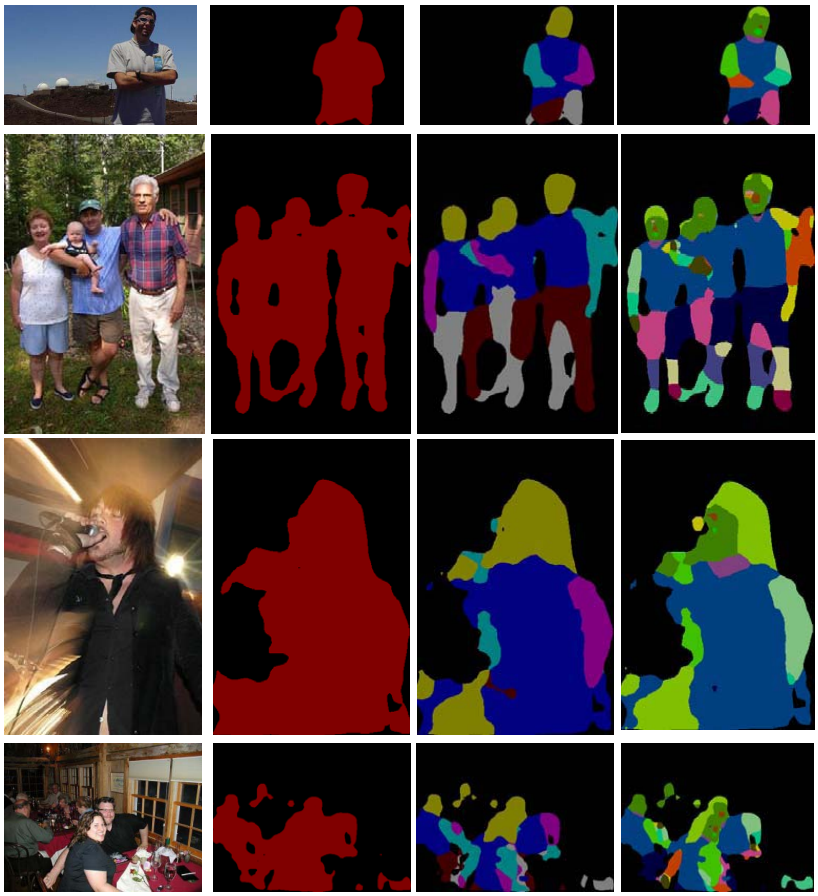

Fig. 4. Exemplar results of transductive learning. Column 1 shows the input images. Columns 2–4 show the predicted maps for binary labels (human and others), five part labels (torso, head, neck, arm, and leg), and the other subpart labels, respectively.

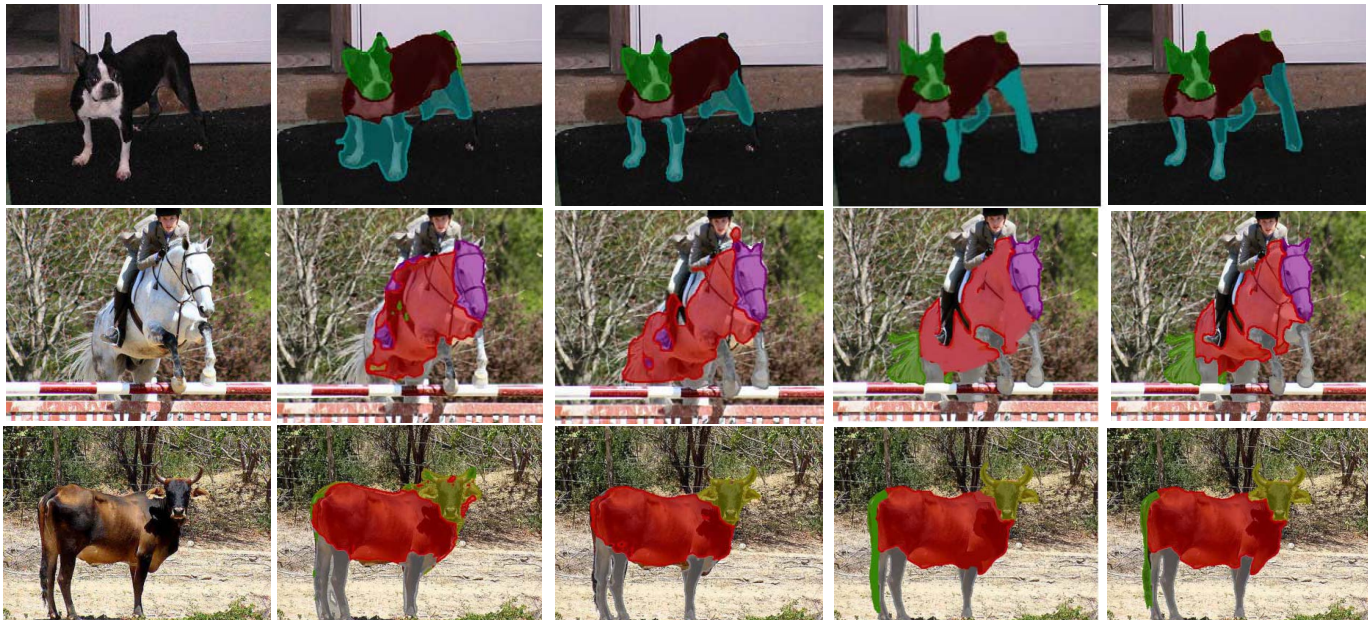

Fig. 5. Exemplar results of part parsing. Column 1: input images. Column 2: results by [19]. Column 3: results by [43]. Column 4: results by the proposed method (Our-I). Column 5: ground-truth label map. Color codes are randomly generated to highlight the segmented regions.

For each image in the first column, we visualize its label map using three figures for clarity. Column 1: human and background, Column 2: torso, head, neck, arm, and leg, and 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 are used for training.

C. Results on the Quadrupeds Data Set [43]

We further test and evaluate the proposed method over the Quadrupeds data set. We used the same baseline methods as the previous experiment. Table II 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 categorywise IOU. For every method, we also average categorywise 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 semisupervised 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 multilabel 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 includes 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. Conclusion

This article presented a multilabel FCN that can be effectively trained on semisupervised 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: 1) pairwise coherences between part labels and their offspring labels; 2) pixelwise manifold regularization; and 3) exclusive constraints between object parts labels. We formulated these objectives in a unified loss function and use it to learn deep features in the semisupervised setting. The proposed model can be end-to-end trained using the standard stochastic gradient algorithm.

Experiments with comparisons on public image data sets showed that our method can achieve state-of-the results for segmenting object parts of varying semantic levels in images. We empirically showed that: 1) additional use of a large amount of unlabeled images brought significant improvements in multilevel part segmentation; 2) our method achieved comparable performance while using 50% or less labeled samples than the alternatives; and 3) 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 semisupervised hierarchical object parsing.

References

[1] M. R. Amer, D. Xie, M. Zhao, S. Todorovic, and S.-C. Zhu, "Cost-sensitive top-down/bottom-up inference for multiscale activity recognition," in *Proc. Eur. Conf. Comput. Vis.* Springer, 2012, pp. 187–200.

[2] M. Belkin and P. Niyogi, "Laplacian eigenmaps and spectral techniques for embedding and clustering," in *Advances in Neural Information Processing System*, vol. 14. Cambridge, MA, USA: MIT Press, 2002, pp. 585–591.

[3] M. Belkin, P. Niyogi, and V. Sindhwani, "Manifold regularization: A geometric framework for learning from labeled and unlabeled examples," *J. Mach. Learn. Res.*, vol. 7, pp. 2399–2434, Nov. 2006.

[4] Y. Bengio, A. Courville, and P. Vincent, "Representation learning: A review and new perspectives," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 8, pp. 1798–1828, Aug. 2013.

[5] L. Bourdev, S. Maji, T. Brox, and J. Malik, "Detecting people using mutually consistent poselet activations," in *Proc. Eur. Conf. Comput. Vis.* Berlin, Germany: Springer, 2010, pp. 168–181.

[6] L.-C. Chen, S. Fidler, A. L. Yuille, and R. Urtasun, "Beat the MTurkers: Automatic image labeling from weak 3D supervision," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 3198–3205.

[7] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "Semantic image segmentation with deep convolutional nets and fully connected CRFs," 2014, *arXiv:1412.7062*. [Online]. Available: https://arxiv.org/abs/1412.7062

[8] X. Chen, R. Mottaghi, X. Liu, S. Fidler, R. Urtasun, and A. Yuille, "Detect what you can: Detecting and representing objects using holistic models and body parts," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2014, pp. 1971–1978.

[9] J. Dai, K. He, Y. Li, S. Ren, and J. Sun, "Instance-sensitive fully convolutional networks," in *Proc. Eur. Conf. Comput. Vis.* Amsterdam, The Netherlands: Springer, 2016, pp. 534–549.

[10] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, Jun. 2005, vol. 1, no. 1, pp. 886–893.

[11] P. Duygulu, K. Barnard, J. F. G. de Freitas, and D. A. Forsyth, "Object recognition as machine translation: Learning a lexicon for a fixed image vocabulary," in *Proc Eur. Conf. Comput. Vis.* Copenhagen, Denmark: Springer, 2002, pp. 97–112.

[12] L. Fei-Fei, R. Fergus, and P. Perona, "One-shot learning of object categories," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 28, no. 4, pp. 594–611, Apr. 2006.

[13] P. Felzenszwalb, D. McAllester, and D. Ramanan, "A discriminatively trained, multiscale, deformable part model," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2008, pp. 1–8.

[14] R. Girshick, F. Iandola, T. Darrell, and J. Malik, "Deformable part models are convolutional neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 437–446.

[15] R. Girshick, J. Shotton, P. Kohli, A. Criminisi, and A. Fitzgibbon, "Efficient regression of general-activity human poses from depth images," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Nov. 2011, pp. 415–422.

[16] M. Guillaumin, D. Küttel, and V. Ferrari, "ImageNet auto-annotation with segmentation propagation," *Int. J. Comput. Vis.*, vol. 110, no. 3, pp. 328–348, 2014.

[17] M. Guillaumin, T. Mensink, J. Verbeek, and C. Schmid, "TagProp: Discriminative metric learning in nearest neighbor models for image auto-annotation," in *Proc. IEEE 12th Int. Conf. Comput. Vis.*, Sep./Oct. 2009, pp. 309–316.

[18] F. Han and S.-C. Zhu, "Bottom-up/top-down image parsing by attribute graph grammar," in *Proc. 10th IEEE Int. Conf. Comput. Vis. (ICCV)*, vol. 2, Oct. 2005, pp. 1778–1785.

[19] B. Hariharan, P. Arbeláez, R. Girshick, and J. Malik, "Hypercolumns for object segmentation and fine-grained localization," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 447–456.
[20] P. Krähenbühl and V. Koltun, "Efficient inference in fully connected CRFs with Gaussian edge potentials," in *Proc. Adv. Neural Inf. Process. Syst*, 2011, vol. 2, no. 3, p. 4.
[21] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2012, pp. 1097–1105.
[22] J. B. Kruskal, "Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis," *Psychometrika*, vol. 29, no. 1, pp. 1–27, Mar. 1964.
[23] B. Kulis, S. Basu, I. Dhillon, and R. Mooney, "Semi-supervised graph clustering: A kernel approach," *Mach. Learn.*, vol. 74, no. 1, pp. 1–22, Jan. 2009.
[24] M. P. Kumar, H. Turki, D. Preston, and D. Koller, "Learning specific-class segmentation from diverse data," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Nov. 2011, pp. 1800–1807.
[25] V. S. Lempitsky, P. Kohli, C. Rother, and T. Sharp, "Image segmentation with a bounding box prior," in *Proc. IEEE 12th Int. Conf. Comput. Vis.*, Jul. 2009, pp. 277–284.
[26] T.-Y. Lin *et al.*, "Microsoft COCO: Common objects in context," in *Proc. Eur. Conf. Comput. Vis.* Zürich, Switzerland: Springer, 2014, pp. 740–755.
[27] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 3431–3440.
[28] W.-L. Lu, J.-A. Ting, J. J. Little, and K. P. Murphy, "Learning to track and identify players from broadcast sports videos," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 7, pp. 1704–1716, Jul. 2013.
[29] F. Nie, D. Xu, I. W. Tsang, and C. Zhang, "Flexible manifold embedding: A framework for semi-supervised and unsupervised dimension reduction," *IEEE Trans. Image Process.*, vol. 19, no. 7, pp. 1921–1932, Jul. 2010.
[30] Z.-Y. Niu, D.-H. Ji, and C. L. Tan, "Word sense disambiguation using label propagation based semi-supervised learning," in *Proc. 43rd Annu. Meeting Assoc. Comput. Linguistics*, 2005, pp. 395–402.
[31] G. Papandreou, L.-C. Chen, K. P. Murphy, and A. L. Yuille, "Weakly- and semi-supervised learning of a deep convolutional network for semantic image segmentation," in *Proc. IEEE Int. Conf. Comput. Vis.*, Jun. 2015, pp. 1742–1750.
[32] D. Pathak, E. Shelhamer, J. Long, and T. Darrell, "Fully convolutional multi-class multiple instance learning," 2014, *arXiv:1412.7144*. [Online]. Available: https://arxiv.org/abs/1412.7144
[33] M. Pei, Y. Jia, and S.-C. Zhu, "Parsing video events with goal inference and intent prediction," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Nov. 2011, pp. 487–494.
[34] P. O. Pinheiro and R. Collobert, "From image-level to pixel-level labeling with convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 1713–1721.
[35] P. O. Pinheiro, R. Collobert, and P. Dollar, "Learning to segment object candidates," in *Proc. Adv. Neural Inf. Process. Syst.*, 2015, pp. 1990–1998.
[36] C. Rother, V. Kolmogorov, and A. Blake, "grabcut: Interactive foreground extraction using iterated graph cuts," *ACM Trans. Graph.*, vol. 23, no. 3, pp. 309–314, 2004.
[37] X. Song, T. Wu, Y. Jia, and S.-C. Zhu, "Discriminatively trained and-or tree models for object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2013, pp. 3278–3285.
[38] J. B. Tenenbaum, V. de Silva, and J. C. Langford, "A global geometric framework for nonlinear dimensionality reduction," *Science*, vol. 290, no. 5500, pp. 2319–2323, Dec. 2000.
[39] Z. Tu, X. Chen, A. L. Yuille, and S.-C. Zhu, "Image parsing: Unifying segmentation, detection, and recognition," *Int. J. Comput. Vis.*, vol. 63, no. 2, pp. 113–140, 2005.
[40] J. Verbeek and B. Triggs, "Region classification with Markov field aspect models," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2007, pp. 1–8.
[41] A. Vezhnevets, V. Ferrari, and J. M. Buhmann, "Weakly supervised structured output learning for semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2012, pp. 845–852.
[42] O. Vinyals, C. Blundell, T. Lillicrap, K. Kavukcuoglu, and D. Wierstra, "Matching networks for one shot learning," in *Proc. Adv. Neural Inf. Process. Syst.*, 2016, pp. 3630–3638.
[43] P. Wang, X. Shen, Z. Lin, S. Cohen, B. Price, and A. Yuille, "Joint object and part segmentation using deep learned potentials," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2015, pp. 1573–1581.
[44] J. Weston, F. Ratle, H. Mobahi, and R. Collobert, "Deep learning via semi-supervised embedding," in *Neural Networks: Tricks of the Trade*. Berlin, Germany: Springer, 2012, pp. 639–655.
[45] W. Xia, C. Domokos, J. Dong, L.-F. Cheong, and S. Yan, "Semantic segmentation without annotating segments," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2013, pp. 2176–2183.
[46] R. Mottaghi *et al.*, "The role of context for object detection and semantic segmentation in the wild," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2014, pp. 891–898.
[47] J. Xu, A. G. Schwing, and R. Urtasun, "Learning to segment under various forms of weak supervision," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 3781–3790.
[48] S. Yan, D. Xu, B. Zhang, H.-J. Zhang, Q. Yang, and S. Lin, "Graph embedding and extensions: A general framework for dimensionality reduction," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 29, no. 1, pp. 40–51, Jan. 2007.
[49] J. Zhu, J. Mao, and A. L. Yuille, "Learning from weakly supervised data by the expectation loss SVM (e-SVM) algorithm," in *Proc. Adv. Neural Inf. Process. Syst.*, 2014, pp. 1125–1133.
[50] X. J. Zhu, "Semi-supervised learning literature survey," Dept. Comput. Sci., Univ. Wisconsin-Madison, Madison, WI, USA, 2005.

Xiaobai Liu received the Ph.D. degree from the Huazhong University of Science and Technology, Wuhan, China, in 2012.

He is currently an Associate Professor of computer science with San Diego State University (SDSU), San Diego, CA, USA. He has authored or coauthored more than 50 peer-reviewed articles in top-tier conferences (e.g., ICCV and CVPR) and leading journals [e.g., the IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE (TPAMI) and the IEEE TRANSACTIONS ON IMAGE PROCESSING (TIP)]. His current research interests include scene parsing with a variety of topics, e.g., joint inference for recognition and reconstruction, commonsense reasoning, and so on.

Dr. Liu was a recipient of a number of awards for his academic contribution, including the 2013 Outstanding Thesis Award by China Computer Federation (CCF).

Qian Xu received the B.S. degree from the School of Science, Beihang University, Beijing, China, in 2006, and the master's and Ph.D. degrees from the Department of Statistics, San Diego State University, San Diego, CA, USA, in 2011 and 2017, respectively.

She is currently the Co-Founder and President of XreLab Inc., San Diego. Her current research interests include various statistical models and their applications in computer vision.

Grayson Adkins received the master's degree from the Department of Computer Science, San Diego State University, San Diego, CA, USA.

His current research interests include computer vision and machine learning.

Eric Medwedeff is currently pursuing the Ph.D. degree with San Diego State University, San Diego, CA, USA, and the University of California at Irvine, Irvine, CA, under a Joint Doctoral Program.

His current research interests include computer vision and machine learning.

Liang Lin (M'09–SM'15) was a Postdoctoral Fellow with the University of California at Los Angeles, Los Angeles, CA, USA, from 2008 to 2010. From 2014 to 2015, he was a Senior Visiting Scholar with The Hong Kong Polytechnic University, Hong Kong, and The Chinese University of Hong Kong, Hong Kong. He is currently a Full Professor of Sun Yat-sen University, Guangzhou, China. He has authored and coauthored more than 100 articles in top-tier academic journals and conferences (e.g., 10 papers in the IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE (TPAMI)/*International Journal of Computer Vision* (IJCV) and more than 40 articles in CVPR/ICCV/NIPS/IJCAI).

Prof. Lin is a fellow of IET. He was a recipient of the Best Paper Runner-Up Award in ACM NPAR 2010, the Google Faculty Award in 2012, the Best Paper Diamond Award at IEEE ICME in 2017, and the Hong Kong Scholars Award in 2014. He served as an Area/Session Chair for numerous conferences such as ICME, ACCV, and ICMR. He has been serving as an Associate Editor for the IEEE TRANSACTIONS ON HUMAN-MACHINE SYSTEMS, *The Visual Computer*, and *Neurocomputing*.

Shuicheng Yan is currently a Chief Scientist with Qihoo 360 Inc., China, and also the Dean's Chair Associate Professor with the National University of Singapore, Singapore. He has authored/coauthored hundreds of technical articles over a wide range of research topics, with Google Scholar citation over 20 000 times and h-index (66). He is an ISI Highly Cited Researcher from 2014 to 2016. His current research interests include machine learning, computer vision, and multimedia.

Prof. Yan is a fellow of IAPR. He was a recipient of Winner or Honorable-Mention Prizes (seven times) in PASCAL VOC and ILSVRC competitions with his team and more than ten times best (student) paper prizes.