End-to-end Learning for Graph Decomposition

Jie Song[1] Bjoern Andres[3,4] Michael J. Black[2] Otmar Hilliges[1] Siyu Tang[1,2,4]
[1] ETH Zurich [2] MPI for Intelligent Systems [3] Bosch Center for AI [4] University of Tübingen

Abstract

Deep neural networks provide powerful tools for pattern recognition, while classical graph algorithms are widely used to solve combinatorial problems. In computer vision, many tasks combine elements of both pattern recognition and graph reasoning. In this paper, we study how to connect deep networks with graph decomposition into an end-to-end trainable framework. More specifically, the minimum cost multicut problem is first converted to an unconstrained binary cubic formulation where cycle consistency constraints are incorporated into the objective function. The new optimization problem can be viewed as a Conditional Random Field (CRF) in which the random variables are associated with the binary edge labels. Cycle constraints are introduced into the CRF as high-order potentials. A standard Convolutional Neural Network (CNN) provides the front-end features for the fully differentiable CRF. The parameters of both parts are optimized in an end-to-end manner. The efficacy of the proposed learning algorithm is demonstrated via experiments on clustering MNIST images and on the challenging task of real-world multi-people pose estimation.

1. Introduction

Many computer vision problems such as multi-person pose estimation [35], instance segmentation [21], and multi-target tracking [42] can be viewed as optimization problems, where the decompositions of a graph are feasible solutions. For example, in multi-person pose estimation, a graph $G = (V, E)$ can be constructed where the nodes V correspond to body joint detections and the edges E connect a person's joints [35]. A partitioning of the graph G into connected components that correspond to the joints of a single individual can be found, for example, via solving the Minimum Cost Multicut Problem [4, 8].

This formulation has several appealing properties: First, it does not favor one decomposition over another and the number of graph components is determined by the solution in an unbiased fashion. Contrary to this, some balanced cut problems [39] rely on a fixed number of graph components or introduction of biases into the problem definition. Second, it is straightforward to utilize this optimization problem in practice: for many vision tasks, an input graph can be easily constructed and the cost of the incident nodes belonging to distinct components can be obtained robustly using Deep Neural Networks, e.g. [14, 21].

By far the most common way of applying the minimum cost multicut problem to vision tasks is to employ a multi-stage pipeline [15, 21, 35, 43]. First, the task dependent detections and the affinity measures between the detections are obtained by two separately trained networks. Second, the coefficients of the objective function are constructed based on the output of the networks and third, the optimization is performed independently on top of the detection graph by either branch and bound algorithms [35, 42] or heuristic greedy search algorithms [6].

While straightforward, a notable downside of this multi-stage approach is that the deep networks are learned locally. That is, the dependencies among the optimization variables are not considered during the training of the deep feature representations. However, it has been shown that graphical models such as Conditional Random Fields (CRFs) can increase the performance of deep feature learning approaches [44, 50]. In this work we then ask the question whether the global dependencies defined by a general graph decomposition problem, such as the minimum cost multicut problem, can lead to learning of better feature representations.

Motivated by this question, we propose an end-to-end trainable framework to learn feature representations globally in a graph decomposition problem. We first convert the minimum cost multicut problem into an unconstrained binary cubic problem to incorporate the hard consistency constraints into the objective function. The appealing property of this new optimization problem is that it can be viewed as a conditional random field (CRF). The random variables of the CRF are associated with the binary edge labels of the initial graph, and the hard constraints can be introduced as high-order potentials in the CRF. We further propose an end-to-end learnable framework that consists of a standard Convolutional Neural Network (CNN) as the front-end and a fully differentiable CRF with the high-order potentials.

The advantages of the proposed framework are: (i) The

parameters of the CRF and the weights of the front-end CNN are optimized jointly during the training of the full network via backpropagation. This joint training facilitates a learnable balance between the unary potentials and high-order potentials that enforce the validity of the edge labeling, which leads to a better decomposition. (ii) The cycle inequalities, encoded by the high-order potentials, serve as supervision signals during learning of the deep feature representations. This meta-supervision from the global consistency constraints is complementary to the direct local supervision (standard CNN training). In this way it teaches the network how to behave by taking the dependencies among the output random variables into account.

In experiments, we first present analyses on the task of clustering MNIST ([24]) images, showing that the proposed method improves the feature learning via enforcing the global consistency constraints. Then the applicability of the proposed approach on the challenging task of multi-person pose estimation is demonstrated. Results suggest the effectiveness of the end-to-end learning framework in terms of better feature learning, cycle constraint validity, tighter confidence of the marginal estimates and final pose estimation performance.

2. Related Work

The minimum cost multicut problem. The multicut problem has been explored for various computer vision tasks [14, 19, 25, 35, 42, 21]. In [15, 35], a joint node and edge labeling problem is proposed to model the multi-person pose estimation task. In [42, 43], the multi-target tracking task is formulated as a graph decomposition problem. Meanwhile, many algorithms for efficiently solving the minimum cost multicut problem have been developed [5, 17, 18, 20, 31, 48, 41]. Beier at al. [5] propose a correlation clustering fusion method which iteratively improves the current solution by a fusion operation. The proposed algorithm maintains a valid decomposition at all times. Yarkony et al. [48] relies on column generation to combine feasible solutions of subproblems into successively better solutions in planar graphs. Swoboda and Andres [41] propose a dual decomposition and linear program relaxation algorithm.

There are also algorithms that integrate optimization problems as layers into network architectures for end-to-end training [1, 11, 37, 49]. Schulter et al. [37] propose a joint learning framework for the cost functions of network flow problems. Amos and Kolter [1] develop a general method for integrating quadratic programs with deep networks. Due to the cubic complexity in the number of constraints, it is an open question whether this method can be applied to complex vision tasks. Funke et al. [12] propose to use a structured loss for training an instance segmentation network with an iterative region agglomeration algorithm for the task of neuron segmentation from electron microscopy. To the best of our knowledge, ours is the first work to introduce an end-to-end learnable framework for the multicut formulation by reformulating of the cycle constraints as high-order terms in a CRF model.

Learning deep structured model. Several approaches propose to jointly learn the feature representations and the structural dependency between the variables of interest [3, 7, 9, 27, 40]. Chen et al. [7] propose a learning framework to estimate the deep representations and the parameters of their Markov random field model together. Zheng et al. [50] reformulate the mean field iterations for the CRFs as recurrent neural network layers with Gaussian pairwise potentials. The front-end CNNs and the recurrent neural network can be trained end-to-end with the usual back-propagation algorithm. Arnab et al. [3] extend the model proposed in [50] by incorporating object detection and superpixel information as high-order potentials for the task of image semantic segmentation. Chu et al. [9] propose a model to implicitly incorporate the structural information into the hidden feature layers of their CNN. The goal of our work is to design an end-to-end learnable framework for the minimum cost multicut problem. While the mean field inference used here does not guarantee a feasible graph decomposition, it effectively allows integration of CNN and graph decomposition.

Human pose estimation. Recent deep neural network methods have made great progress on human pose estimation in natural images in particular for the single person case [26, 28, 30, 34, 44, 47]. As for a more general case where multiple people are present in images, previous work can mainly be grouped into either top-down or bottom-up categories. Top-down approaches first detect individual people and then predict each person's pose [10, 13, 33]. The top-down approaches generally achieve better performance on public benchmarks, because they can leverage external powerful person detection models, turning the pose estimation task into the simpler single-person case. Bottom-up approaches directly detect individual body joints and then associate them with individual people [6, 14, 15, 27, 32, 46, 38]. In [6, 35], the body joint detections and the affinity measures between the detections are first trained by deep networks, then the association is performed independently either by branch and bound algorithms [35] or by heuristic greedy search algorithms [6]. One potential advantage over top-down approaches is that the decision making for detections (typically non-maximum suppression is deployed) is performed at lower levels (joints) rather than at the highest level (person). Note that in [27], the associations are trained by predicting person IDs alongside the joint detections.In contrast, our method focuses on end-to-end learning of the graph decomposition problem.

3. Optimization Problems

3.1. Minimum Cost Multicut Problem

The minimum cost multicut problem [4, 8] is a constrained binary linear program with respect to a graph $G = (V, E)$ and a cost function $c : E \to \mathbb{R}$:

$$\min_{y \in \{0,1\}^E} \quad \sum_{e \in E} c_e \, y_e \tag{1}$$

$$\text{subject to} \quad \forall C \in \text{cc}(G) \, \forall e \in C : \quad y_e \leq \sum_{e' \in C \setminus \{e\}} y_{e'} \ . \tag{2}$$

Here, the optimization variables $y \in \{0, 1\}^E$ correspond to a binary labeling of the edges E. $y_e = 1$ indicates that the edge e is cut. In other words, the nodes v and w connected by edge e are in distinct components of G. $\text{cc}(G)$ denotes the set of all chord-less cycles of G. The cycle constraints in Eq. 2 define the feasible edge labelings, which relate one-to-one to the decompositions of the graph G.

A toy example is illustrated in Fig. 1: (a) shows an example graph G; (b) is a valid decomposition of G; and (c) shows an invalid solution that violates the cycle inequalities (Eq. 2). The cost function $c : E \to \mathbb{R}$ is characterized by model parameters θ. In previous work [14, 15, 35], the cost function is defined as $\log \frac{1-p_e}{p_e}$, where p_e denotes the probability of y_e being cut. Given a feature f_e on the edge e, p_e takes a logistic form: $\frac{1}{1+\exp(-\langle \theta, f_e \rangle)}$. The maximal probable model parameters θ are then obtained by maximum likelihood estimation on training data. f_e can be attained via some deep feature representations extracted from a separately trained deep network. For example, in [14] and [43], f_e is obtained from a CNN and a Siamese network respectively.

Research questions. At the heart of this work lie the following research questions: first, how to jointly optimize the model parameters θ and the weights of the underlying deep neural network for the graph decomposition problem? Second, how to utilize the cycle consistency constraints as a supervision signal and to capture the dependencies between the output random variables during training? In the following, we present our end-to-end learnable framework, which provides solutions to these research questions.

3.2. Unconstrained Binary Cubic Problem

Our first observation is that the minimum cost multicut problem can be equivalently stated as an unconstrained binary multilinear program with a large enough $K \in \mathbb{N}$

$$\min_{y \in \{0,1\}^E} \quad \sum_{e \in E} c_e \, y_e + K \sum_{C \in \text{cc}(G)} \sum_{e \in C} y_e \prod_{e' \in C \setminus \{e\}} (1 - y_{e'}) \ . \tag{3}$$

In the special case where G is complete, every 3-cycle is chordless. Thus, Eq. 3 specializes to the binary *cubic* problem, where $\bar{y}_{vw} := 1 - y_{vw}$:

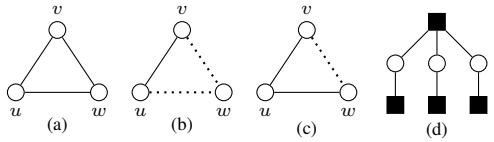

Figure 1: We illustrate a graph G in (a); a feasible solution and an infeasible solution are shown in (b) and (c) respectively; the factor graph of the CRF model of G is in (d).

$$\min_{y \in \{0,1\}^E} \quad \sum_{e \in E} c_e \, y_e + K \sum_{\{u,v,w\} \in \binom{V}{3}} (y_{uv} \bar{y}_{vw} \bar{y}_{uw} + \bar{y}_{uv} y_{vw} \bar{y}_{uw} + \bar{y}_{uv} \bar{y}_{vw} y_{uw}) \ . \tag{4}$$

An invalid cycle labeling, e.g. Fig. 1(c) where $y_{vw} = 1, y_{uw} = y_{uv} = 0$ and $\bar{y}_{uv} y_{vw} \bar{y}_{uw} = 1$, contributes a value K into the objective (Eq. 4). By setting K to be large enough, the right-hand side terms in Eq. 4 are forced to be 0, so that the cycle consistent constraints defined in Eq. 2 are satisfied.

3.3. Multicut as Conditional Random Fields

Our second observation is that the unconstrained binary cubic problem (Eq. 4) can be expressed by a Conditional Random Field (CRF) with unary potentials that are defined on each edge variable and high-order potentials that are defined on every three edge variables. More specifically, we define a random field over the variables $\mathbf{X} = (X_1, X_2 \cdots, X_{|E|})$ that we want to predict. $\mathbf{I}$ is the observation, in our case this is an image. The optimization problem (Eq. 4) can be expressed as the following CRF model:

$$E(\mathbf{x}|\mathbf{I}) = \sum_i \psi_i^U(x_i) + \sum_c \psi_c^{Cycle}(\mathbf{x}_c) \tag{5}$$

where we associate each random variable x_i with an edge variable y_e in Eq. 4. The random variable x_i takes a value from the label set $\{0, 1\}$. Furthermore, each $\mathbf{x}_c$ in Eq. 5 is associated with every three edge variables, namely y_{uv}, y_{vw} and y_{uw}, where $\{u, v, w\} \in \binom{V}{3}$. $E(\mathbf{x}|\mathbf{I})$ is the energy associated with a configuration $\mathbf{x}$ conditioned on the observation $\mathbf{I}$. Our goal is to obtain a labeling with minimal energy, namely $\hat{\mathbf{x}} \in \text{argmin}_{\mathbf{x}} E(\mathbf{x}|\mathbf{I})$. Such a labeling is the maximum a posteriori (MAP) solution of the Gibbs distribution $P(\mathbf{X} = \mathbf{x}|\mathbf{I}) = \frac{1}{Z(\mathbf{I})} \exp\left(-E(\mathbf{x}|\mathbf{I})\right)$ defined by the energy $E(\mathbf{x}|\mathbf{I})$, where $Z(\mathbf{I})$ is the partition function.

Unary potentials. The unary potentials $\psi_i^U(x_i)$ correspond to the left-hand side term in Eq. 4, measuring the inverse likelihood of an edge being cut or not. The unary potential can take inputs from various sources. As shown in Sec. 4, in case of multi-person pose estimation, $\psi_i^U(x_i)$ can directly be the output of a state-of-the-art CNN [6].

High-order potentials. The high-order terms $\psi_c^{Cycle}(\mathbf{x}_c)$ are introduced to model the cycle inequalities (Eq. 2) in the minimum cost multicut problem and correspond to the right-hand side terms in Eq. 4. Each high-order potential associates a cost to a cycle in the initial graph. The primary idea is that, for every cycle in the graph, a high cost is incurred if the current edge labelings in the cycle violate the consistency constraint. More specifically, for a fully connected graph, each cycle in the graph consists of three edges. There are three types of valid edge labelings (1-1-0,1-1-1,0-0-0) and one type of invalid edge labeling (0-0-1) that violates the constraints defined in Eq. 2. Fig. 1 illustrates a simple graph and examples of valid (1-1-0) and invalid (1-0-0) edge labelings. To assign high/low cost for the invalid/valid cycles, we deploy the *pattern-based potentials* proposed in [23]

$$\psi_c^{Cycle}(\mathbf{x}_c) = \begin{cases} \gamma_{\mathbf{x}_c} & \text{if } \mathbf{x}_c \in \mathcal{P}_c \\ \gamma_{max} & \text{otherwise,} \end{cases} \tag{6}$$

where $\mathcal{P}_c$ is the set of recognized label configurations for the clique, namely, valid cycles in the initial graph. We assign a cost $\gamma_{\mathbf{x}_c}$ to each of them. γ_{max} is then assigned to all the invalid label configurations for the clique, namely, invalid cycles in the initial graph.

Given the proposed potentials, minimizing the energy of the proposed CRF model (Eq. 5) is then equivalent to minimizing the optimization problem defined in Eq. 4.

Inference. We resort to the mean-field update formulation of [50] to minimize the energy defined in Eq. 5 iteratively as part of the joint framework. For the mean field inference, an alternative distribution $Q(\mathbf{x})$ defined over the random variables is introduced to minimize the KL-divergence between $Q(\mathbf{x})$ and the true distribution $P(\mathbf{x})$. The general mean field update follows [22]:

$$Q_i(x_i = l) = \frac{1}{Z_i} \exp\{-\sum_{c \in C} \sum_{\{\mathbf{x}_c | x_i = l\}} Q_{c-i}(\mathbf{x}_{c-i}) \psi_c(\mathbf{x}_c)\}. \tag{7}$$

Here $\mathbf{x}_c$ is a configuration of all the variables in the clique c and $\mathbf{x}_{c-i}$ is a configuration of all the variables in the clique c except x_i. Given the definition of the pattern-based potential in Eq. 6, the mean field updates for our CRF model can be derived from the work of [45] as:

$$\begin{aligned} Q_i^t(x_i = l) = \frac{1}{Z_i} \exp\{-\sum_{c \in C} (\sum_{p \in \mathcal{P}_{c|x_i=l}} (\prod_{j \in c, j \neq i} Q_j^{t-1}(x_j = p_j)) \gamma_p \\ + \gamma_{max} (1 - (\sum_{p \in \mathcal{P}_{c|x_i=l}} (\prod_{j \in c, j \neq i} Q_j^{t-1}(x_j = p_j)))))\} \end{aligned} \tag{8}$$

where x_j represents a random variable in the clique c apart from x_i, $\mathcal{P}_{c|x_i=l}$ is the subset of $\mathcal{P}_c$ where $x_i = l$. t denotes the t^{th} iteration of the mean field inference. Assume L is the value of a loss function defined on the result obtained by the mean filed inference, Eq. 8 allows us to backpropagate the error $\frac{\partial L}{\partial Q}$ to the input $\mathbf{x}$ and the parameters $\gamma_{\mathbf{x}_c}$, γ_{max}.

Note that the mean field inference does not guarantee that we obtain a valid graph decomposition. In our formulation the inference enforces the validity of the cycle consistency but does not guarantee that all the hard constraints (Eq. 2) are fulfilled. Therefore in practice, we resort to fast heuristics (e.g. [19]) to return a feasible graph decomposition following the mean field inference.

Learning. Leveraging the message passing update (Eq. 8) allows us to backpropagate the error signals, which facilitates the whole learning mechanism. More specifically, we are now able to jointly optimize the deep feature representation and the parameters for partitioning of the graph, by reformulating the original optimization problem into a CRF model. The following parameters can be jointly learned via backpropagation:

- W: the weights of the front-end neural network
- θ: characterizing the cost function $c : E \rightarrow \mathbb{R}$
- $\gamma_{\mathbf{x}_c}$, γ_{max}: parameters of the high-order potentials.

By joint training, the dependencies between the optimization variables are incorporated into the learning for a better deep feature representation via the proposed high-order potentials.

3.4. Example: Clustering MNIST Digits

To understand how the proposed end-to-end learning model integrates the dependencies among the output random variables during training, we consider a simple task that clusters images of hand-written digits (MNIST [24]) *without* specifying the number of clusters. This problem can be formulated as a minimum cost multicut problem (Eq. 1-2) that is defined on a fully connected graph. The nodes of the graph indicate the digits and edges connect the images that hypothetically indicate the same digit. Through this simple task, we discuss two approaches to learn feature representations used to associate the images.

Approach I: Standalone Siamese network. A straight forward way to obtain the similarity measures between any two images is to train a Siamese network which takes a pair of images as input and produces a probabilistic estimation to indicate whether they are the same or different digits. We use the architecture of LeNet [24] which is commonly used on digit classification tasks. Fig. 2 shows two example results. In Fig. 2 (a), the probabilities for the top/left pair and left/right pair being the same digit are 0.96 and 0.86 respectively, which are correctly estimated. But for the top/right pair, it is 0.48, likely due to the high intra-class variation. Similarly for the example in Fig. 2 (b), the probability for

(a) (b)

Figure 2: Examples of inconsistent edge labels produced by a stand alone Siamese network on the MNIST digits.

the top/right pair being the same digit is incorrectly estimated. When we partition these digits into clusters, the incorrect similarity estimates introduce invalid cycles. Now the question is whether we can deploy cycle constraints to learn a better Siamase network, resulting in more robust and consistent similarity measures.

Approach II: Train CRF and Siamese network jointly. In this approach, we aim to train the Siamese network by taking the cycle consistency constraints into account. For this we leverage our formulation, where the partitioning problem is converted into the energy minimization problem defined in the CRF (Eq. 5). Specifically, we add a stack of customized inference layers that perform the iterative mean field updates with high-order potentials on top of the Siamese network (the details are in Sec. 4.2). Now we are able to train the Siamese network and the CRF model jointly. In this configuration the probability of the top/right pair (in Fig 2 (a)) indicating the same digit is increased to 0.57 (+0.09), using the end-to-end learned Siamese network. It is further improved to 0.62 (+0.14) after the mean-field updates with the jointly learned CRF parameters. In terms of overall performance, the accuracy of similarity measures is increased from 91.5% to 93.2%. The corresponding final clustering accuracy is increased from 94.1% to 95.9%.

Discussion: This simple setting illustrates that our approach can produce more robust and consistent results on clustering task such as the MNIST digits. The next open question is how to design a jointly learnable framework for more challenging real-world vision tasks that rely on clustering.

4. Multi-person Pose Estimation

In this section, we further design an end-to-end learnable framework for the task of multi-person pose estimation.

Our network consists of four parts: 1) a front-end CNN that outputs feature representations (Sec. 4.1); 2) fully connected layers to convert the features to the unary potentials (Sec. 4.1); 3) a stack of customized layers that perform the iterative mean field updates (Sec. 4.2) and 4) the loss layer on top of the mean field iteration (Sec. 4.3). We choose multi-person pose estimation as case study because this task is considered to be one of the fundamental problems in understanding people in natural images. Recent work [13, 35, 6, 14] has made significant progress on this task. For instance, the work proposed by Cao et al. [6] presents a powerful deep neural network to learn feature representation for body joints and limbs, followed by a fast heuristic matching algorithm to associate body joints to individual poses. Given the performance of [6] on benchmarks, in the following, we utilize their network architecture as the front-end CNN. Our approach is complementary to [6] in that our focus is the joint optimization of the deep feature learning and the detection association.

4.1. From CNN to Unary Potentials

Network Architecture. The network proposed in [6] has two separate branches after sharing the same basic convolutional layers: one branch predicts the confidence maps for 14 body joints and the other branch estimates a set of part affinity fields, which encode joint to joint relations. The part field is a 2D vector field. More specifically, each pixel in the affinity field is associated with an estimated 2D vector that encodes the direction pointing from one joint to the other. In [6], the part fields are implemented only for pairs of joints that follow the kinematic tree of the human body, e.g. left elbow to left hand. However, in order to incorporate high-order potentials among neighboring joints, we train the model to also capture the feature between non-adjacent detections, e.g. shoulder to wrist.

Graph Construction. Given an input image, we first obtain the body joint candidates from the detection confidence maps. For each type of joint, we keep multiple detection hypotheses even for those that are in close proximity. A detection graph is constructed by introducing edges for pairs of hypotheses that describe the same type of body joint, and for pairs of hypotheses between two different joints. Note that the constructed graph is not fully connected but every chordless cycle in the graph consists of only three edges.

Edge Feature. The key to the robust graph decomposition is a reliable feature representation on the edges to indicate whether the corresponding joint detections belong to the same/different person. For the edges that connect the detection hypotheses of different body types, we use the corresponding part field estimation. More specifically, we compute the inner product between the unit vector defined by the direction of the edge and vectors that are estimated by the part field. We collect 10 values by uniformly sampling along the line segment defined by the edge. These values form the feature f_e for the corresponding edge. For the edges that connect the detection hypotheses of the same joint type, we simply use the euclidean distance between the detection as the feature.

The Unary ψ^U. It is straightforward to construct the unary potentials $\psi_i^U(x_i)$ (Eq. 5) from the edge feature f_e. We incorporate two fully connected layers to encode the feature to classify if an edge is cut, namely, the two corresponding joints belong to different persons. As described in Sec. 3.3, during training, we can obtain the error signal from the mean

field updates to learn the parameters of the fully connected layers and the front-end CNN that produces the edge feature.

4.2. Mean Field Updates

Zheng et al. [50] propose to formulate the mean field iteration as recurrent neural network layers, and [3] further extend it to include high-order object detection and superpixel potentials for the task of semantic segmentation. In this work, we follow their framework with the modification of incorporating the proposed pattern-based potentials. The goal of the mean field iterations is to update the marginal distribution $Q_i^t(x_i = l)$. For initialization, $Q_i^1(x_i = l) = \frac{1}{Z_i}\exp\{-\psi_i^U(x_i = l)\}$ is performed, where $Z_i = \sum_l \exp\{-\psi_i^U(x_i = l)\}$. This is equivalent to applying a soft-max function over the negative unary energy across all the possible labels for each link. This operation does not include any parameters and the error can be back-propagated to the front-end convolutional or fully connected layers where the unary potentials come from. Once the marginal has been initialized, we compute the high-order potentials based on Eq. 8. Specifically, the valid cliques in $\mathcal{P}_c$ are 0-0-0, 1-1-1 and 1-1-0, while the non-valid cliques are 0-0-1, where 1 indicates that the corresponding edge is cut. This operation is differentiable with respect to the parameters $\gamma_{\mathbf{x}_c}$ and γ_{max} introduced in Eq. 8, allowing us to optimize them via backpropagation. The errors can also flow back to $Q^1(X)$. Once the high-order potential is obtained, it is summed up with the unary potential and then the sum is normalized via the soft-max function to generate the new marginal for the next iteration. Multiple mean-field iterations can be efficiently implemented by stacking this basic operation. During the inference, as the mean field inference does not guarantee a feasible solution to the original optimization problem, we use the fast heuristic proposed in [6] as an additional step to come back to the feasible set.

4.3. Loss and Training

During training, we first train the joint confidence maps and part affinity field maps with a standard $L2$ loss as described in [6]. Once the basic features are learned, the next step is to train the unary with the softmax loss function. This is performed in an on-the-fly manner, which means the detection hypotheses for the body joints are estimated and then the links between the hypotheses are also established during training time. Their ground-truth labels are also generated online at the same time. The final step is to train the parameters of the CRF with high-order potentials with a softmax loss function in an end-to-end manner along with the basic convolutional and fully connected layers.

4.4. Experiments

Dataset. We use the MPII Human Pose dataset [2] which consists of about 25k images and contains around 40k total annotated people. The training and test split contain 3844 and 1758 groups of people respectively. We conduct ablation experiments on a held out validation set. During testing, no information about the number of people or the height of individuals is provided. We deploy the evaluation metric proposed by [35] as our final association measure. The metric is calculated as the average precision of the joint detections for all the people in the images. In the following experiments, we use shortcuts for body joints (Head-H, Neck-N, Shoulder-S, Elbow-E, Wrist-W, Hip-Hi, Knee-K, Ankle-A).

	H-N	N-S	S-E	E-W	S-Hi	Hi-K	K-A	Mean
origin	0.755	0.656	0.662	0.558	0.679	0.593	0.611	0.635
Iter 1	0.792	0.699	0.696	0.591	0.718	0.631	0.644	0.663
Iter 2	0.811	0.716	0.719	0.613	0.731	0.649	0.656	0.675
Iter 3	0.819	0.721	0.725	0.617	0.735	0.654	0.662	0.685

Table 1: **Marginal distribution updates.** Numbers represent evolution of the marginal probabilities along with the mean-field iterations for different type of limbs.

	H-N-S	S-E-W	N-LH-RH	H-K-A	Mean
origin	1.68	3.40	1.41	3.83	2.60
Iter 1	1.08	2.63	1.01	3.05	2.02
Iter 2	0.98	2.48	0.88	2.76	1.78
Iter 3	0.91	2.42	0.85	2.68	1.67

Table 2: **Ratio of non valid cycle.** Numbers (%) represent the ratio of non valid cycle for four different types of cliques that are defined for adjacent body joints.

Implementation Details. The front-end CNN architecture has several stacked fully convolutional layers with an input size of 368x368 (cf. [6]). We train the basic CNN using a batch size of 12 with a learning rate of 1e-4. For training of the CRF parameters, the learning rate is 1e-5. The whole architecture is implemented in Caffe [16]. As for runtime efficiency, on the validation set, the mean field inference takes about $0.3ms$ and the whole inference time of the proposed end-to-end framework is around $88ms$ on average.

Effectiveness of the CRF Inference. To demonstrate the effectiveness of our proposed mean-field layers approximating the CRF inference, we evaluate the evolution of the marginal distribution for the random variables X_i. In the case of pose estimation, each X_i in the CRF represents a link between two body joints. Tab. 1 summarizes 7 different types of such links. Each row shows the average marginal probabilities $Pr(X = 0)$ for the links with ground-truth of 0, where the label 0 indicates that the edge should not be cut. The marginal probability can be read as the confidence that two joints belong to the same person. The marginal distributions for all limbs in Tab. 1 increase with each iteration even for very challenging combinations, e.g. Elbow-Wrist and Knee-Ankle. After three iterations of inference, the update converges. We use this setting for further experiments.

Validity of the Cycle Constraints. Another important measurement is the ratio of non-valid cycles after the mean field iterations. Recall that the type of non-valid 3-clique is link-link-cut (Sec. 4). Tab. 2 shows that, with CRF inference, the ratio of non-valid cycles decreases, indicating the effectiveness of the high-order potential.

Benefit of End-to-End Learning on feature representation. One of the key motivations to train the CNN and CRF jointly is to obtain better feature representations. We illustrate this via inspection of the part field feature maps before and after the CRF inference. Fig. 3 shows that the confidence maps generally increase in sharpness and contain less noise. This is particularly apparent for images that contain heavy occlusions; e.g. in the second image in the second row, the limbs of the partially occluded people become more distinguishable, suggesting a notable improvement in the feature learning, especially for the challenging cases (see highlights in blue). This confirms one of our central assumptions, motivating this work. The learned features are more informative if learned with additional supervision signals from the high-order terms.

Return to a Feasible Solution. After the inference, we do not obtain a valid graph decomposition directly. Some heuristics (either the greedy search [6] or the KL heuristic [14]) are required to generate a valid decomposition efficiently. We evaluate these two heuristics with three different settings: 1) only front-end CNN and fully connected layers (unary); 2) separately trained CRF and front-end CNN (unary and CRF); 3) end-to-end training of the whole network (end-to-end). Tab. 3 summarizes the respective performance on the validation set. The advantage of the end-to-end strategy over the baselines is clearly observable. Fig. 4 shows that these improvements are more pronounced in the most challenging cases with heavy occlusion, where modeling the high-order dependencies among the variables has the most impact.

Comparison With Others. We compare ours with other methods on the MPII Human Pose dataset. Tab. 4 summarizes the results. Note that, as described in Sec. 2, there are in general two types of approaches for multi-person pose estimation: bottom-up approaches and top-down approaches. On public benchmarks, the top-down approaches generally achieve better performance because they can leverage external powerful person detection models, turning the pose estimation task into the simpler single-person case. In contrast, bottom-up approaches first detect joint candidates and then cluster them into individual skeletons. In this work, we focus on improving performance in the bottom-up setting since it is a direct match for the proposed end-to-end learnable graph decomposition method.

Our implementation of Cao et al., [6] serves as baseline and achieves 75.2 mAP, whereas our end-to-end method increases this accuracy to 76.7 mAP. Given that the available dataset is relatively small and bottom-up methods seem to have saturated, this improvement is notable.

Method	Head	Shou	Elbo	Wris	Hip	Knee	Ankl	Mean
unary (KL)	88.55	83.98	71.43	60.97	73.44	65.25	56.66	71.32
unary and CRF (KL)	89.26	84.57	72.34	61.65	73.93	66.98	58.32	72.15
end-to-end (KL)	89.52	85.13	72.92	62.41	74.43	67.33	58.75	72.96
unary (greedy)	91.30	86.14	73.69	62.84	73.40	66.43	58.73	73.21
unary and CRF (greedy)	91.43	86.93	74.96	64.71	74.12	67.36	59.97	74.39
end-to-end (greedy)	**91.70**	**87.48**	**75.43**	**65.23**	**74.57**	**67.99**	**60.61**	**75.02**

Table 3: **Ablation study** on the validation set. End-to-end training notably increases accuracy of multi-person pose estimation.

Method	Head	Shou	Elbo	Wris	Hip	Knee	Ankl	Mean
Bottom-up methods:								
Insafutdinov et al., [15]	78.4	72.5	60.2	51.0	57.2	52.0	45.4	59.5
Pishchulin et al., [35]	89.4	84.5	70.4	59.3	68.9	62.7	54.6	70.0
Insafutdinov et al., [14]	88.8	87.0	75.9	64.9	74.2	68.8	60.5	74.3
Cao et al., [6]	91.2	87.6	77.7	66.8	75.4	68.9	61.7	75.6
Our Baseline	90.7	87.4	77.3	66.5	75.7	69.0	60.9	75.2
Our Method with CRF	**91.8**	**88.3**	**78.5**	**67.8**	**77.1**	**70.0**	**63.0**	**76.7**
Top-down methods (use separate person detector or single-person pose refinement):								
Fang et al., [10]	88.4	86.5	78.6	70.4	74.4	73.0	65.8	76.7
Newell et al., [27]	92.1	89.3	78.9	69.8	76.2	71.6	64.7	77.5
Nie et al., [29]	92.2	89.7	82.1	74.4	78.6	76.4	69.3	**80.4**

Table 4: **Comparison on MPII Human Pose** dataset. Ours outperforms all other bottom-up methods by a good margin and is comparable to top-down methods. Top-down methods can leverage larger datasets to train external person detectors.

The bottom-rows in Tab. 4 also list methods that utilize a person detector or single-person pose refinement. Specifically, the method in [27] uses a single-person pose estimator to refine the final result, and [10] uses a separate Faster R-CNN [36] person detector. [29] proposes a hybrid model which combines the top-down and bottom-up information.

5. Conclusion, Limitation and Future Work

In this work, our goal is to answer the following research questions: (1) how to jointly optimize the model parameters and the weights of the underlying deep neural network for the graph decomposition problem? (2) how to use the cycle consistency as a supervision signal to capture the dependencies of the output random variables during training? To that end, we propose to convert the minimum cost multicut problem to an energy minimization problem defined in a CRF. The hard constraints of the multicut problem are formulated as high-order potentials of the CRF whose parameters are learnable. We perform analyses on the task of clustering digit images and multi-person pose estimation. The results validate the potential of our method and show improvement both for the feature learning and the final clustering task.

Although, as we show in this work, the proposed learning method for the multicut problem has several strong points, there are still some limitations. First, with the proposed mean field update, we can jointly learn the front end deep networks and the parameters of the graph decomposition. However, the hard constraints in the optimization problem are not guaranteed to be satisfied. Therefore during testing, we resort

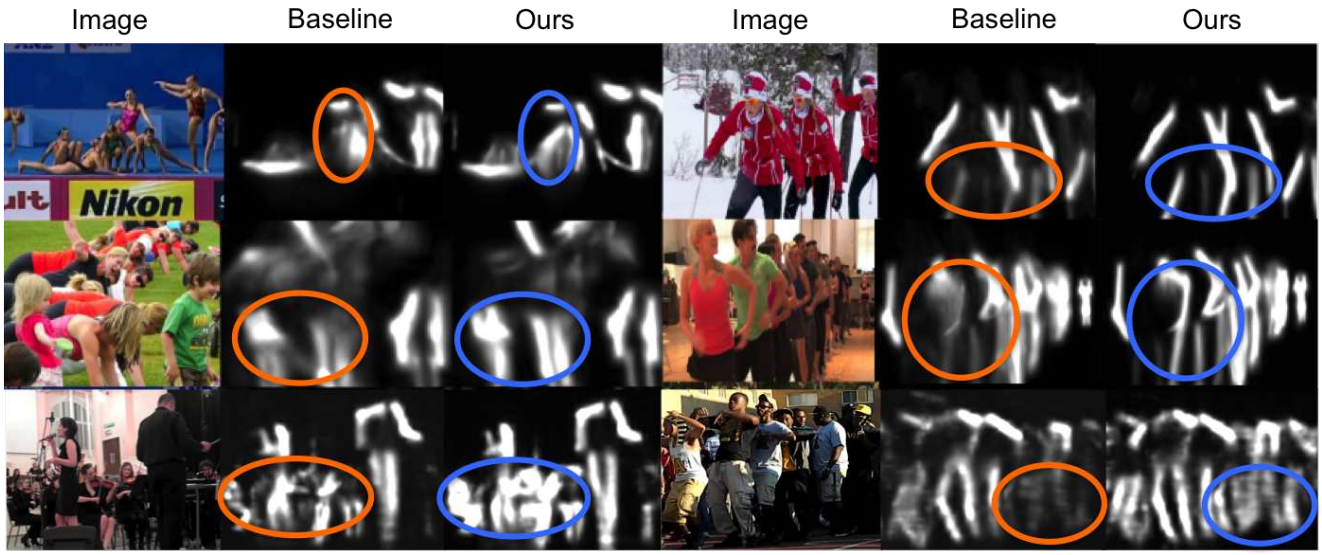

Figure 3: **Feature learning comparison.** Left: input image; Middle: part field map learned locally; Right: part field map learned with the cycle consistency. The right samples clearly show sharper and more accurate confidence maps.

Figure 4: **Qualitative Results.** Left: association without CRF; Right: association after inference. **First row**, obvious wrong connections are corrected by inference. In the **second row** occluded people are separated. The last example is a failure case.

to efficient heuristic solvers to return a feasible graph decomposition. Second, we show notable improvement on the feature learning and validity of the cycle inequality for the multi-person pose estimation task, but the final performance gain on pose association does not support us to outperform the state-of-the-art top-down methods. One reason is that our end-to-end training only operates on the part affinity field, not on the body joint detections, which is crucial for the final result. To include the body joint detections in the end-to-end training pipeline is a practical future direction. Nevertheless, We think that this work adds an important primitive to the toolbox of the graph decomposition problem and opens up many avenues for future research.

Acknowledgement. We thank Nvidia for the donation of GPUs used in this work. S. Tang acknowledges funding by Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) Projektnummer 276693517 SFB 1233.
Disclosure. MJB has received research gift funds from Intel, Nvidia, Adobe, Facebook, and Amazon. While MJB is a part-time employee of Amazon, his research was performed solely at, and funded solely by, MPI. MJB has financial interests in Amazon and Meshcapade GmbH.

References

[1] Brandon Amos and J. Zico Kolter. OptNet: Differentiable optimization as a layer in neural networks. In *the International Conference on Machine Learning (ICML)*, 2017. 2

[2] Mykhaylo Andriluka, Leonid Pishchulin, Peter Gehler, and Bernt Schiele. 2d human pose estimation: New benchmark and state of the art analysis. In *the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014. 6

[3] Anurag Arnab, Sadeep Jayasumana, Shuai Zheng, and Philip H. S. Torr. Higher order conditional random fields in deep neural networks. In *the European Conference on Computer Vision (ECCV)*, 2016. 2, 6

[4] Nikhil Bansal, Avrim Blum, and Shuchi Chawla. Correlation clustering. *Machine Learning*, 56(1-3):89–113, 2004. 1, 3

[5] Thorsten Beier, Fred A Hamprecht, and Jorg H Kappes. Fusion moves for correlation clustering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015. 2

[6] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. In *the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 1, 2, 3, 5, 6, 7

[7] Liang-Chieh Chen, Alexander G. Schwing, Alan L. Yuille, and Raquel Urtasun. Learning deep structured models. In *the International Conference on Machine Learning (ICML)*, 2015. 2

[8] Sunil Chopra and Mendu R Rao. The partition problem. *Mathematical Programming*, 59(1-3):87–115, 1993. 1, 3

[9] Xiao Chu, Wanli Ouyang, and Xiaogang Wang. CRF-CNN: Modeling structured information in human pose estimation. In *Advances in Neural Information Processing Systems (NIPS)*, 2016. 2

[10] Haoshu Fang, Shuqin Xie, Yu-Wing Tai, and Cewu Lu. Rmpe: Regional multi-person pose estimation. In *The IEEE International Conference on Computer Vision (ICCV)*, 2017. 2, 7

[11] Davi Frossard and Raquel Urtasun. End-to-end learning of multi-sensor 3d tracking by detection. In *the IEEE International Conference on Robotics and Automation, ICRA*, 2018. 2

[12] Jan Funke, Fabian Tschopp, William Grisaitis, Arlo Sheridan, Chandan Singh, Stephan Saalfeld, and Srinivas C Turaga. Large scale image segmentation with structured loss based deep learning for connectome reconstruction. *IEEE transactions on pattern analysis and machine intelligence*, 41(7):1669–1680, 2018. 2

[13] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *The IEEE International Conference on Computer Vision (ICCV)*, 2017. 2, 5

[14] Eldar Insafutdinov, Mykhaylo Andriluka, Leonid Pishchulin, Siyu Tang, Evgeny Levinkov, Bjoern Andres, and Bernt Schiele. ArtTrack: Articulated multi-person tracking in the wild. In *the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 1, 2, 3, 5, 7

[15] Eldar Insafutdinov, Leonid Pishchulin, Bjoern Andres, Mykhaylo Andriluka, and Bernt Schiele. Deepercut: A deeper, stronger, and faster multi-person pose estimation model. In *the European Conference on Computer Vision (ECCV)*, 2016. 1, 2, 3, 7

[16] Yangqing Jia, Evan Shelhamer, Jeff Donahue, Sergey Karayev, Jonathan Long, Ross Girshick, Sergio Guadarrama, and Trevor Darrell. Caffe: Convolutional architecture for fast feature embedding. In *the ACM international conference on Multimedia*, 2014. 6

[17] Jörg Hendrik Kappes, Markus Speth, Björn Andres, Gerhard Reinelt, and Christoph Schn. Globally optimal image partitioning by multicuts. In *the International Workshop on Energy Minimization Methods in Computer Vision and Pattern Recognition*, 2011. 2

[18] Jörg Hendrik Kappes, Markus Speth, Gerhard Reinelt, and Christoph Schnörr. Higher-order segmentation via multicuts. *Comput. Vis. Image Underst.*, 143(C):104–119, Feb. 2016. 2

[19] Margret Keuper, Evgeny Levinkov, Nicolas Bonneel, Guillaume Lavoué, Thomas Brox, and Bjoern Andres. Efficient decomposition of image and mesh graphs by lifted multicuts. In *the IEEE International Conference on Computer Vision (ICCV)*, 2015. 2, 4

[20] Sungwoong Kim, Sebastian Nowozin, Pushmeet Kohli, and Chang D Yoo. Higher-order correlation clustering for image segmentation. In *Advances in neural information processing systems (NIPS)*, pages 1530–1538, 2011. 2

[21] Alexander Kirillov, Evgeny Levinkov, Bjoern Andres, Bogdan Savchynskyy, and Carsten Rother. InstanceCut: from edges to instances with multicut. In *the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 1, 2

[22] Daphne Koller and Nir Friedman. *Probabilistic Graphical Models: Principles and Techniques - Adaptive Computation and Machine Learning*. The MIT Press, 2009. 4

[23] Nikos Komodakis and Nikos Paragios. Beyond pairwise energies: Efficient optimization for higher-order mrfs. In *the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009. 4

[24] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998. 2, 4

[25] Evgeny Levinkov, Alexander Kirillov, and Bjoern Andres. A comparative study of local search algorithms for correlation clustering. In *German Conference on Pattern Recognition (GCPR)*, 2017. 2

[26] Yue Luo, Jimmy Ren, Zhouxia Wang, Wenxiu Sun, Jinshan Pan, Jianbo Liu, Jiahao Pang, and Liang Lin. Lstm pose machines. In *the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 2

[27] Alejandro Newell, Zhiao Huang, and Jia Deng. Associative embedding: End-to-end learning for joint detection and grouping. In *Advances in Neural Information Processing Systems (NIPS)*, 2017. 2, 7

[28] Alejandro Newell, Kaiyu Yang, and Jia Deng. Stacked hourglass networks for human pose estimation. In *the European Conference on Computer Vision (ECCV)*, 2016. 2

[29] Xuecheng Nie, Jiashi Feng, Junliang Xing, and Shuicheng Yan. Pose partition networks for multi-person pose estimation.

In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018. 7

[30] Xuecheng Nie, Jiashi Feng, Yiming Zuo, and Shuicheng Yan. Human pose estimation with parsing induced learner. In *the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 2

[31] Sebastian Nowozin and Stefanie Jegelka. Solution stability in linear programming relaxations: Graph partitioning and unsupervised learning. In *the International Conference on Machine Learning (ICML)*, 2009. 2

[32] George Papandreou, Tyler Zhu, Liang-Chieh Chen, Spyros Gidaris, Jonathan Tompson, and Kevin Murphy. Personlab: Person pose estimation and instance segmentation with a bottom-up, part-based, geometric embedding model. In *the European Conference on Computer Vision (ECCV)*, 2018. 2

[33] George Papandreou, Tyler Zhu, Nori Kanazawa, Alexander Toshev, Jonathan Tompson, Chris Bregler, and Kevin Murphy. Towards accurate multiperson pose estimation in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 2

[34] Xi Peng, Zhiqiang Tang, Fei Yang, Rogerio S. Feris, and Dimitris Metaxas. Jointly optimize data augmentation and network training: Adversarial data augmentation in human pose estimation. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 2

[35] Leonid Pishchulin, Eldar Insafutdinov, Siyu Tang, Bjoern Andres, Mykhaylo Andriluka, Peter V Gehler, and Bernt Schiele. Deepcut: Joint subset partition and labeling for multi person pose estimation. In *the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 1, 2, 3, 5, 6, 7

[36] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in Neural Information Processing Systems (NIPS)*, 2015. 7

[37] Samuel Schulter, Paul Vernaza, Wongun Choi, and Manmohan Krishna Chandraker. Deep network flow for multi-object tracking. *the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 2

[38] Taiki Sekii. Pose proposal networks. In *the European Conference on Computer Vision (ECCV)*, 2018. 2

[39] Jianbo Shi and Jitendra Malik. Normalized cuts and image segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.*, 22(8):888–905, Aug. 2000. 1

[40] Jie Song, Limin Wang, Luc Van Gool, and Otmar Hilliges. Thin-slicing network: A deep structured model for pose estimation in videos. In *the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 2

[41] Paul Swoboda and Bjoern Andres. A message passing algorithm for the minimum cost multicut problem. In *the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 2

[42] Siyu Tang, Bjoern Andres, Miykhaylo Andriluka, and Bernt Schiele. Subgraph decomposition for multi-target tracking. In *the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015. 1, 2

[43] Siyu Tang, Mykhaylo Andriluka, Bjoern Andres, and Bernt Schiele. Multiple people tracking by lifted multicut and person re-identification. In *the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 1, 2, 3

[44] Jonathan J Tompson, Arjun Jain, Yann LeCun, and Christoph Bregler. Joint training of a convolutional network and a graphical model for human pose estimation. In *Advances in Neural Information Processing Systems (NIPS)*, 2014. 1, 2

[45] Vibhav Vineet, Jonathan Warrell, and Philip H. S. Torr. Filter-based mean-field inference for random fields with higher-order terms and product label-spaces. *International Journal of Computer Vision*, 110(3):290–307, Dec 2014. 4

[46] Shaofei Wang, Alexander Ihler, Konrad Kording, and Julian Yarkony. Accelerating dynamic programs via nested benders decomposition with application to multi-person pose estimation. In *the European Conference on Computer Vision (ECCV)*, 2018. 2

[47] Shih-En Wei, Varun Ramakrishna, Takeo Kanade, and Yaser Sheikh. Convolutional pose machines. In *the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 2

[48] Julian Yarkony, Alexander Ihler, and Charless C. Fowlkes. Fast planar correlation clustering for image segmentation. In Andrew Fitzgibbon, Svetlana Lazebnik, Pietro Perona, Yoichi Sato, and Cordelia Schmid, editors, *the European Conference on Computer Vision (ECCV)*, 2012. 2

[49] Andrei Zanfir and Cristian Sminchisescu. Deep learning of graph matching. In *the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 2

[50] Shuai Zheng, Sadeep Jayasumana, Bernardino Romera-Paredes, Vibhav Vineet, Zhizhong Su, Dalong Du, Chang Huang, and Philip H. S. Torr. Conditional random fields as recurrent neural networks. In *the IEEE International Conference on Computer Vision (ICCV)*, 2015. 1, 2, 4, 6