# LEARNING GRAPH DECOMPOSITION

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

We propose a novel end-to-end trainable framework for the graph decomposition problem. 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 of the initial graph and the hard constraints are introduced in the CRF as high-order potentials. The parameters of a standard Neural Network and the fully differentiable CRF can be optimized in an end-to-end manner. We demonstrate the proposed learning algorithm in the context of clustering of hand written digits, particularly in a setting where no direct supervision for the graph decomposition task is available, and multiple person pose estimation from images in the wild. The experiments validate the effectiveness of our approach both for the feature learning and for the final clustering task.

## 1 INTRODUCTION

Many computer vision problems, e.g. multi-person pose estimation (Pishchulin et al., 2016), instance segmentation (Kirillov et al., 2017), and multi-target tracking (Tang et al., 2015), can be viewed as optimization problems, where decompositions of a graph are the 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 the detections that hypothetically indicate the same person (Pishchulin et al., 2016). Partitioning the detections that describe the same person into the same connected component with respect to the graph $G$ is a Minimum Cost Multicut Problem (Chopra & Rao, 1993; Bansal et al., 2004), with respect to a linear objective function. It has several appealing properties: First, in contrast to other balanced cut problems (Shi & Malik, 2000), it does not favor one decomposition over another. Instead of relying on a fixed number of graph components or biasing them by the problem definition, in this formulation the number of decompositions is determined by the solution in an unbiased fashion. 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 cutting edges connecting incident nodes can be obtained robustly from some Deep Neural Network, e.g. Insafutdinov et al. (2017); Kirillov et al. (2017).

By far, the most common way of applying the minimum cost multicut problem to vision tasks is to employ a multi-stage pipeline (Kirillov et al., 2017; Pishchulin et al., 2016; Tang et al., 2015; Keuper et al., 2015b;a; Insafutdinov et al., 2016; Tang et al., 2017; Insafutdinov et al., 2017). In case of multi-person pose estimation: First, the joint 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 (Pishchulin et al., 2016) or heuristic greedy search algorithms (Cao et al., 2017).

While this approach is straightforward, it is noteworthy that the deep networks are learned independently without utilizing the knowledge of how to perform the graph decomposition globally and dependencies among the optimization variables are not considered during learning of the deep feature representations. Furthermore, because many datasets are relatively small, such networks often do not generalize well to different test datasets. This in turn implies that the optimization-based graph decomposition itself is not leveraged to its full potential due to the already accurate unary estimates. Thus the question arises whether a more tightly linked interplay between representa-

tion learning and graph decomposition can improve robustness and generalization of deep learning approaches in tasks that rely on clustering.

Motivated by this question, we propose a novel end-to-end trainable framework for the joint learning of feature representation and graph decomposition problem. Since discrete optimization problems are generally not differentiable we first convert the minimum cost multicut problem to an unconstrained binary cubic problem. The appealing property of this new optimization problem is that it can be viewed as a conditional random field (CRF) with hardconstraints being represented by high-order potentials. Furthermore, the CRF is fully differentiable and can be integrated with a CNN based front-end for feature learning into an end-to-end trainable network.

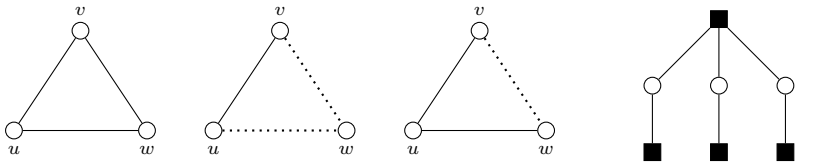
The advantages of the proposed framework are: 1) the proposed CRF model facilitates a learnable balance between the unary potentials and high-order potentials that enforce the validity of the edge labeling, which leads to a better decomposition. 2) the cycle inequalities, encoded by the high-order potentials, enforce global consistency constraints during learning of the deep feature representations. Importantly, these constraints are complementary to the direct local supervision (standard CNN training) in the sense that it explicitly teaches the network to learn correlations between the output random variables. 3) in case of the absence of the direct local supervision, the validity of the cycle inequalities serves as supervisory signals to train the parameters of the CRF model, providing a weakly-supervised learning mechanism for the graph decomposition problem.

We demonstrate the proposed model on the tasks of clustering images of hand-written digits, in particular in a meta-supervised setting, and multi-person pose estimation from images. Our experimental results suggest the effectiveness of the end-to-end learning framework in terms of cycle constrain validity, tighter confidence of the marginal estimates and for feature representation learning.

**Related Work.** The minimum cost multicut problem has been explored for diverse set of computer vision tasks (Pishchulin et al. (2016); Insafutdinov et al. (2017); Tang et al. (2015); Kirillov et al. (2017); Keuper et al. (2015b); Levinkov et al. (2017)). Keuper et al. (2015b) applies it to the motion segmentation task, where pixel-wise motion trajectories are clustered into individual moving objects. In Pishchulin et al. (2016); Insafutdinov et al. (2016), a joint node and edge labeling problem is proposed to model the multi-person pose estimation task. In Tang et al. (2015; 2017), the multi-target tracking task is formulated as a graph decomposition problem. To the best of our knowledge, ours is the first work that introduces an end-to-end learning framework for the multicut formulation.

Several works have been proposed to jointly learn the feature representations and the structural dependency between the variables of interest (Chen et al., 2015; Arnab et al., 2016; Newell et al., 2017; Chu et al., 2016). Chen et al. (2015) proposes a learning framework to jointly estimate the deep representations and the parameters of their Markov random field model. Zheng et al. (2015) proposes to formulate the mean field iterations as recurrent neural network layers, and Arnab et al. (2016) further extends Zheng et al. (2015) to include their object detection and superpixel potentials for the task of semantic segmentation. Chu et al. (2016) proposes a CRF-CNN model to incorporate the structural information into the hidden feature layers of their CNN.

Recent deep neural network based methods have made great progress on human pose estimation in natural images in particular for the single person case (Tompson et al., 2014; Newell et al., 2016; Wei et al., 2016). 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 persons pose (Fang et al., 2017; Papandreou et al., 2017; He et al., 2017). One of the challenges for top-down approaches is that they make detection decisions at a very early stage, which is fragile and prone to false negatives. Bottom-up approaches directly detect individual body joints and then associate them with individual people (Cao et al., 2017; Insafutdinov et al., 2016; 2017; Newell et al., 2017). In Pishchulin et al. (2016); Cao et al. (2017), 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 (Pishchulin et al., 2016) or by heuristic greedy search algorithms (Cao et al., 2017). One potential advantage over top-down approaches is that the decision making of detections (typically non-maximum suppression is deployed) is performed at lower levels (joints) rather than at the highest level (person).




(a) A graph $G$ (b) A feasible solution (c) Violation of the cycle constraint (d) The graphical model defined on the edge labeling of $G$

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

## 2 Optimization Problem

### 2.1 Minimum Cost Multicut Problem

The minimum cost multicut problem (Chopra & Rao, 1993; Bansal et al., 2004) is a constrained binary linear program w.r.t. a graph $G = (V, E)$ and a cost funciton $c : E \to \mathbb{R}$:

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

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

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 labellings, 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 $\theta$. In previous work (Pishchulin et al., 2016; Insafutdinov et al., 2016; 2017), 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 $\theta$ 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 (Insafutdinov et al., 2017) and (Tang et al., 2017), $f_e$ is obtained from a convolutional neural network and a Siamese network respectively.

At the heart of this work lie the following research questions: first, how to jointly optimize the parameters $\theta$ and the weights of the underlying deep neural network for the graph decomposition problem? Second, how to utilize the cycle consistency constraints as 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 the these research questions.

### 2.2 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 constant $C \in \mathbb{N}$

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

In the special case where $G$ is complete, every 3-cycle is chordless. Thus, Eq. 3 specializes to the binary *cubic* problem as described in Eq. 4 where $\bar{y}_{vw} := 1 - y_{vw}$.

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

### 2.3 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 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, e.g. an image. We associate each random variable $x_i$ with an edge variable $y_e$ in Eq. 4, and the random variable $x_i$ takes a value from a label set $\{0, 1\}$. Then the optimization problem 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 $E(\mathbf{x}|\mathbf{I})$ 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 $\hat{\mathbf{x}} \in \operatorname{argmin}_{\mathbf{x}} E(\mathbf{x}|\mathbf{I})$. Such a labeling is the maximum a posteriori (MAP) solution of the Gibbs distribution $P(\mathbf{X} = \mathbf{x}|\mathbf{I}) = \frac{1}{Z(\mathbf{I})} \exp -E(\mathbf{x}|\mathbf{I})$ defined by the energy $E(\mathbf{x}|\mathbf{I})$, where $Z(\mathbf{I})$ is the partition function.

The unary potentials $\psi_i^U(x_i)$ measure the inverse likelihood of an edge being cut. It can take arbitrary forms. As shown in Sec. 3.2, in case of multi-person pose estimation, $\psi_i^U(x_i)$ utilizes the output of a state-of-the-art CNN (Cao et al., 2017).

The high-order terms $\psi_c^{Cycle}(\mathbf{x}_c)$ are one of the key contributions of this work. They are introduced to model the cycle inequalities (Eq. 2) in the minimum cost multicut problem. 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 will incur if the current edge labellings in the cycle violate the cycle consistency constraint.

**Pattern-based Potentials.** There is a finite set of valid edge labellings for 3-cycles in the graph. Fig. 1 illustrates a simple graph and examples of valid (1-1-0) and invalid (1-0-0) edge labellings. To assign high/low cost for the invalid/valid cycles, we utilize the pattern-based potentials proposed in Komodakis & Paragios (2009).

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

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. $\gamma_{max}$ 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 (5) is then equivalent to minimizing the optimization problem defined in Eq. 4.

**Inference.** We resort to mean field inference to minimize the energy defined in Eq. 5, which has been formulated as a Recurrent Neural Network and integrated into a CNN framework Zheng et al. (2015). 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 Koller & Friedman (2009):

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

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 Vineet et al. (2014) 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 $x_i$, $\mathcal{P}_{c|i=l}$ 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 $\mathbf{x}$ as well as the parameters $\gamma_{\mathbf{x}_c}$ and $\gamma_{max}$. Note that after the mean field inferences, it is not guaranteed to obtain a valid graph decomposition, as the mean field inference enforces the validity of the cycle consistency but does not guarantee that all the hard constraints (E.g. 2) are fulfilled. Therefore in practice, we resort to some fast heuristics (E.g. Keuper et al. (2015b)) to return a feasible graph decomposition after the mean field inferences.

**Learning.** Although the mean field update (Eq. 8) does not guarantee that all the hard constraints (E.g. 2) are fulfilled, it allows us to backpropagate the error signals, which facilitates an end-to-end learning mechanism. More specifically, we are now able to jointly optimize the deep feature representation and the parameters for performing the partitioning of the graph, by reformulating the original optimization problem to the CRF model. Concretely, the following parameters can be jointly learned by the proposed model via backpropagation:

- $W$ which are the weights of the front-end deep neural network
- $\theta$ which characterizes the cost function $c : E \rightarrow \mathbb{R}$ in the minimum cost multicut problem
- $\gamma_{\mathbf{x}_c}$ and $\gamma_{max}$ that are introduced by the high-order potentials of the CRF model.

By the 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. In case that the ground truth labels on the edges are available, e.g. the nodes connected by the edge are in the same/distinct components, we employ the standard cross entropy loss on top of the mean field inferences. Furthermore, in case of the absence of the direct local supervision on the edges, the validity of the cycle inequalities serve as supervision signal to train the parameters of the CRF model, providing a weakly-supervised learning mechanism for the graph decomposition problem.

## 3 APPLICATION

We show applications of the proposed learning algorithm to two tasks: clustering images of hand-written digit (MNIST) (LeCun et al., 1998) and estimation of multi-person poses from images in the wild. In the experiment of clustering MNIST, we illustrate the effectiveness of the learned CRF model for the graph decomposition problem. Particularly in the case of the absence of ground truth labels, we are able to train the CRF by leveraging the validity of cycle inequalities. In the pose estimation experiment, we illustrate the effectiveness of the full end-to-end learning framework with extensive experiments. Our results show that the proposed framework is capable of learning better feature representations, reducing significantly the amount of invalid cycles and of obtaining better final pose estimates.

### 3.1 CLUSTERING IMAGES OF HAND-WRITTEN DIGIT (MNIST)

To validate that the proposed model is an effective approach for the graph decomposition problem, we conduct the following experiment: we partition the MNIST test set into distinct digits, without specifying the number of clusters. This problem can be formulated as a minimum cost multicut problem that is defined on a fully connected graph. The nodes indicate the digit images and edges connect the images that contain the same digit.

**Network Architecture.** Our network to cluster the MNIST consists of four parts: 1) a front end Siamese CNN, built on the architecture of LeNet ( LeCun et al. (1998)) that outputs feature representations for pairs of images; 2) two fully connected layers to convert the features to the unary potentials of the CRF model; 3) a stack of customized layers that perform the iterative mean field updates with high-order potentials; and 4) the loss layer, penalizing invalid cycle inequalities after the mean field iteration.

**Experiment results.** As a baseline, we utilize 1 % of the training set to train a Siamese CNN from scratch to predict, for any pair of images, whether they belong to the same or distinct digits. As shown in Tab. 1, using only 1% of the training data the cut accuracy already reaches 76.1% and

| | 1 % supervision | + 10 % unsuper. | + 30 % unsuper. | + 50 % unsuper. | + 99 % unsuper. |
|---|---|---|---|---|---|
| Acc. Classification | 84.5% | 86.4% | 87.2% | 87.6% | 87.8% |
| Acc. Cut | 76.1% | 78.2% | 79.0% | 79.5% | 79.6% |
| Invalid Cycles | 1.1% | 0.91% | 0.86% | 0.82% | 0.8% |

Table 1: Experiments on the MNIST test dataset.

only 1.1% of all cycles are invalid. After deploying the Kernighan-Lin (KL) heuristic (Insafutdinov et al., 2017) to solve the minimum cost multicut problem on the complete graph, we attain 84.5% for digits clustering (by relating the ten largest clusters to the ten digits optimally, this clustering can be associated with a classification rate)

Next, we add the proposed CRF model on top of the pre-trained Siamese CNN, we utilize the validity of the cycle inequalities as the supervision to train the parameters of the CRF by penalizing the predicted edge labeling in invalid cycles. As shown in Tab. 1, when we add more training images, the accuracy of the edge labels improves steadily, and the percentage of invalid cycles in the graph is reduced constantly. After deploying the KL heuristic on top of the CRF output, the classification accuracy is also improved notably (from $84.5\%$ to $87.8\%$). Note that, the proposed CRF is learned without any ground truth labels on pairs of images containing the same or distinct digits. The only supervision signal is the validity of the edge labeling in the cycles. The empirical results demonstrate the effectiveness of the proposed model for the graph decomposition problem as well as the power of the cycle consistency constraints in the meta-supervised setting.

### 3.2 Multi-person Pose Estimation

In this section, we evaluate our proposed end-to-end learnable algorithm on the challenging task of multi-person pose estimation which is considered to be one of the most fundamental problems in visual understanding of people in natural images. We first briefly introduce the network architecture. Then we demonstrate the effectiveness of the proposed CRF model for updating the marginals based on the high-order potentials. We also measure the ratio of validity for the cycle constraints before and after the mean field inference iterations, showing that the learned mean field inference is able to significantly reduce the amount of invalid cycles. finally we discuss the influence of the end-to-end training on the feature representation learning and the final pose association.

#### 3.2.1 Network Architecture

Similar to the previous section, the network for pose estimation also consists of four parts. The first part is the front end CNN that outputs feature representations. Here, we use the network architecture proposed by Cao et al. (Cao et al., 2017) to effectively learn the deep feature representation for body joints and limbs. The second and the third part are identical to the networks used for the task of clustering MNIST, namely, two fully connected layer to convert the deep feature to the unary potentials of the CRF model and a stack of customized layers that perform the iterative mean field updates with high-order potentials. The last part is a standard cross entropy layer to penalize the predicted edge label if it is inconsistent with the ground truth label. For details of the network structure, please refer to the Appendices.

#### 3.2.2 Experiments

**Dataset.** We use the MPII Human Pose dataset which consists of about 25k images and contains around 40k total annotated people. There is a training and test split with 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 scales of individual is provided. For the final association evaluation, we deploy the evaluation metric proposed by Pishchulin et al. (2016), calculating the average precision of the joint detections for all the people in the images.

**Implementation Details.** The front-end CNN architecture has several stacked fully convolutional layers with an input size of 368x368 as described in Cao et al. (2017). We train the basic CNN using a batch size of 12 with a learning rate of 1e-4. For training the CRF parameters, the learning rate is 1e-5. The whole architecture is implemented in Caffe (Jia et al., 2014).

**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

| | Head-Neck | Neck-Shou | Shou-Elbo | Elbo-Wris | Shou-Hip | Hip-Knee | Knee-Ankl | Mean |
|---|---|---|---|---|---|---|---|---|
| origin | 0.755 | 0.656 | 0.662 | 0.558 | 0.679 | 0.593 | 0.611 | 0.635 |
| Iter 1 | 0.783 | 0.692 | 0.688 | 0.579 | 0.711 | 0.628 | 0.639 | 0.659 |
| Iter 2 | 0.805 | 0.707 | 0.715 | 0.603 | 0.726 | 0.649 | 0.651 | 0.671 |
| Iter 3 | 0.807 | 0.712 | 0.716 | 0.608 | 0.723 | 0.646 | 0.653 | 0.674 |

Table 2: **Marginal distribution updates.** Numbers represent evolution of the marginal probabilities along with the mean-field iterations for different type of limbs.

| | Head-Neck-Shou | Shou-Elbo-Wris | Neck-LHip-RHip | Hip-Knee-Ankl | Mean |
|---|---|---|---|---|---|
| origin | 1.68 | 3.40 | 1.41 | 3.83 | 2.60 |
| Iter 1 | 1.12 | 2.79 | 1.06 | 3.17 | 2.04 |
| Iter 2 | 1.01 | 2.58 | 0.89 | 2.82 | 1.81 |
| Iter 3 | 0.96 | 2.47 | 0.87 | 2.79 | 1.76 |

Table 3: **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.

the random variables $X$. For pose estimation, each variable $X_i$ in the CRF represents a link between two body joints. As seen in Tab. 2, 7 different types of limbs are depicted. The numbers are the average marginal probabilities for those links with the ground truth of not being cut. This index measures how confident a link is supposed to be associated. In other words, the confidence that two joints belong to the same person. As shown in the table, the marginal distributions of all the limbs benefit from high order potentials even for very challenging combinations, e.g. Elbow-Wrist and Knee-Ankle. After three iterations of inference, the update converges and we fix this setting for further experiments.

**Validity of the Cycle Constraints.** Another important measurement for our proposed model is to check the ratio change of non-valid cycles after the mean field iterations. As mentioned in Sec. 3.2, the type of non-valid 3-clique is link-link-cut. We can see from Tab. 3 that, with the CRF inference, the ratios of non-valid cycles decrease, indicating the effectiveness of the high order potential.

**Benefit of End-to-End Learning on Feature Representation.** One of the key advantages of training the CNN and CRF jointly is to obtain a better feature representation. We illustrate it by directly visualizing the part field feature maps before and after the mean field inference. As shown in Fig. 2, the confidence maps in general get sharper and cleaner, particularly for images with 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 for the challenging cases. This is in line with one of the assumptions of this work: the training of the deep features needs additional supervision signals from the high-order terms, particularly for challenging cases.

**Return to a Feasible Solution.** After the CRF inference, we do not obtain a valid graph decomposition directly. Some heuristics (either the greedy search (Cao et al., 2017) or the KL heuristic (Insafutdinov et al., 2017)) are required to generate a valid decomposition efficiently. We evaluate these two heuristics with three different settings for each: 1) only front-end CNN and full connected layers (unary); 2) trained CRF on top of front-end CNN and fully connected layers (unary and CRF); 3) end-to-end finetuning of the whole network ( end-to-end finetuning). Tab. 5 shows the analysis on the validation set and we can see the advantage of the end-to-end strategy over the offline training of


Figure 2: **Feature learning comparison.** Left: input image; Middle: part field map learned locally, without considering the cycle consistency; Right: part field map learned with the cycle consistency. The right samples clearly show sharper and more accurate confidence maps.

Figure 3: **Qualitative Results.** Left: association without CRF inference; Right: association after inference. **First row**, obvious wrong connections are corrected by inference. In the **second row** occluded people are separated. The samples in the **last row** are failure cases.

| 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) | 89.04 | 84.36 | 72.01 | 61.39 | 73.68 | 66.75 | 58.11 | 71.96 |
| end-to-end finetuning (KL) | 89.38 | 84.72 | 72.49 | 61.96 | 74.05 | 66.85 | 58.46 | 72.57 |
| unary (greedy) | 91.30 | 86.14 | 73.69 | 62.84 | 73.40 | 66.43 | 58.73 | 73.21 |
| unary and CRF (greedy) | 91.31 | 86.47 | 74.60 | 64.01 | 73.69 | 66.98 | 59.45 | 73.78 |
| end-to-end finetuning (greedy) | 91.50 | 86.90 | 74.89 | 64.61 | 73.97 | 67.38 | 59.91 | 74.36 |

Table 4: Multi-person pose estimation result on the validation set.

CRF as a post-processing method. As illustrated in Fig. 3, the improvements are mainly achieved on the challenging cases with heavy occlusion, which benefit from modeling the high-order dependency among the variables of interest.

**Comparison With Others.**

We test the proposed method on the MPII Human Pose dataset and compare with other methods. The result are shown in Tab. 5. Our end-to-end method achieves 76.1 mAP, which is on par with other state-of-the-art methods. Note that the method proposed in Newell et al. (2017) uses a single-person pose estimator to refine the final result, and Fang et al. (2017) is a top-down method where a Faster R-CNN Ren et al. (2015) person detector is utilized.

## 4 Conclusion

In this work, we model the minimum cost multicut problem as a CRF. The hard constraints of the multicut problem are formulated as high-order potentials whose parameters are learnable. We further present an end-to-end learning framework for the multi-person pose estimation task. In our framework, the front end CNN and the parameters in the CRF are jointly optimized. The experiments show improvement both for the feature learning and for the final clustering task.

| Method | Head | Shou | Elbo | Wris | Hip | Knee | Ankl | Mean |
|---|---|---|---|---|---|---|---|---|
| Insafutdinov et al., Insafutdinov et al. (2016) | 78.4 | 72.5 | 60.2 | 51.0 | 57.2 | 52.0 | 45.4 | 59.5 |
| pishchulin et al., Pishchulin et al. (2016) | 89.4 | 84.5 | 70.4 | 59.3 | 68.9 | 62.7 | 54.6 | 70.0 |
| Insafutdinov et al., Insafutdinov et al. (2017) | 88.8 | 87.0 | 75.9 | 64.9 | 74.2 | 68.8 | 60.5 | 74.3 |
| Cao et al., Cao et al. (2017) | 91.2 | 87.6 | 77.7 | 66.8 | 75.4 | 68.9 | 61.7 | 75.6 |
| Fang et al., Fang et al. (2017) | 88.4 | 86.5 | 78.6 | 70.4 | 74.4 | 73.0 | 65.8 | 76.7 |
| Newell et al., Newell et al. (2017) | 92.1 | 89.3 | 78.9 | 69.8 | 76.2 | 71.6 | 64.7 | 77.5 |
| Our Method | 91.4 | 87.8 | 78.0 | 67.2 | 76.5 | 69.3 | 62.2 | 76.1 |

Table 5: Comparison with the state-of-the-art on the MPII Human Pose dataset.

## REFERENCES


Anurag Arnab, Sadeep Jayasumana, Shuai Zheng, and Philip H. S. Torr. Higher order conditional random fields in deep neural networks. In *ECCV*, 2016.

Nikhil Bansal, Avrim Blum, and Shuchi Chawla. Correlation clustering. *Machine Learning*, 2004. ISSN 1573-0565. doi: 10.1023/B:MACH.0000033116.57574.95.

Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. In *CVPR*, volume 1, pp. 7, 2017.

Liang-Chieh Chen, Alexander G. Schwing, Alan L. Yuille, and Raquel Urtasun. Learning deep structured models. In *Proceedings of the 32Nd International Conference on International Conference on Machine Learning - Volume 37*, ICML'15, pp. 1785–1794. JMLR.org, 2015. URL http://dl.acm.org/citation.cfm?id=3045118.3045308.

Sunil Chopra and M. R. Rao. The partition problem. *Math. Program.*, 59(1):87–115, March 1993. ISSN 0025-5610. doi: 10.1007/BF01581239. URL http://dx.doi.org/10.1007/BF01581239.

Xiao Chu, Wanli Ouyang, Xiaogang Wang, et al. Crf-cnn: Modeling structured information in human pose estimation. In *Advances in Neural Information Processing Systems*, pp. 316–324, 2016.

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)*, volume 2, 2017.

Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Computer Vision (ICCV), 2017 IEEE International Conference on*, pp. 2980–2988. IEEE, 2017.

Eldar Insafutdinov, Leonid Pishchulin, Bjoern Andres, Mykhaylo Andriluka, and Bernt Schiele. Deepercut: A deeper, stronger, and faster multi-person pose estimation model. In *ECCV*, 2016. doi: 10.1007/978-3-319-46466-4_3.

Eldar Insafutdinov, Mykhaylo Andriluka, Leonid Pishchulin, Siyu Tang, Evgeny Levinkov, Bjoern Andres, and Bernt Schiele. ArtTrack: Articulated multi-person tracking in the wild. In *CVPR*, 2017.

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 *Proceedings of the 22nd ACM international conference on Multimedia*, pp. 675–678. ACM, 2014.

Margret Keuper, Bjoern Andres, and Thomas Brox. Motion trajectory segmentation via minimum cost multicuts. In *ICCV*, 2015a. doi: 10.1109/ICCV.2015.374.

Margret Keuper, Evgeny Levinkov, Nicolas Bonneel, Guillaume Lavoué, Thomas Brox, and Bjoern Andres. Efficient decomposition of image and mesh graphs by lifted multicuts. In *ICCV*, 2015b. doi: 10.1109/ICCV.2015.204.

Alexander Kirillov, Evgeny Levinkov, Bjoern Andres, Bogdan Savchynskyy, and Carsten Rother. InstanceCut: from edges to instances with multicut. In *CVPR*, 2017.

Daphne Koller and Nir Friedman. *Probabilistic Graphical Models: Principles and Techniques - Adaptive Computation and Machine Learning*. The MIT Press, 2009. ISBN 0262013193, 9780262013192.

N. Komodakis and N. Paragios. Beyond pairwise energies: Efficient optimization for higher-order mrfs. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, 2009.

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.

Evgeny Levinkov, Alexander Kirillov, and Bjoern Andres. A comparative study of local search algorithms for correlation clustering. In *German Conference on Pattern Recognition*, pp. 103–114. Springer, 2017.

Alejandro Newell, Kaiyu Yang, and Jia Deng. Stacked hourglass networks for human pose estimation. In *European Conference on Computer Vision*, pp. 483–499. Springer, 2016.

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*, pp. 2274–2284, 2017.

George Papandreou, Tyler Zhu, Nori Kanazawa, Alexander Toshev, Jonathan Tompson, Chris Bregler, and Kevin Murphy. Towards accurate multiperson pose estimation in the wild. *arXiv preprint arXiv:1701.01779*, 8, 2017.

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 *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4929–4937, 2016.

Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 1*, NIPS'15, pp. 91–99, Cambridge, MA, USA, 2015. MIT Press. URL http://dl.acm.org/citation.cfm?id=2969239.2969250.

Jianbo Shi and Jitendra Malik. Normalized cuts and image segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.*, 22(8):888–905, August 2000. ISSN 0162-8828. doi: 10.1109/34.868688. URL https://doi.org/10.1109/34.868688.

Siyu Tang, Bjoern Andres, Miykhaylo Andriluka, and Bernt Schiele. Subgraph decomposition for multi-target tracking. In *CVPR*, 2015. doi: 10.1109/CVPR.2015.7299138.

Siyu Tang, Mykhaylo Andriluka, Bjoern Andres, and Bernt Schiele. Multiple people tracking by lifted multicut and person re-identification. In *CVPR*, 2017.

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*, pp. 1799–1807, 2014.

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. ISSN 1573-1405. doi: 10.1007/s11263-014-0708-6.

Shih-En Wei, Varun Ramakrishna, Takeo Kanade, and Yaser Sheikh. Convolutional pose machines. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4724–4732, 2016.

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 *Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV)*, ICCV '15, pp. 1529–1537, Washington, DC, USA, 2015. IEEE Computer Society. ISBN 978-1-4673-8391-2. doi: 10.1109/ICCV.2015.179. URL http://dx.doi.org/10.1109/ICCV.2015.179.

# A APPENDIX

## A.1 NETWORK ARCHITECTURE FOR MULTI-PERSON POSE ESTIMATION

In this section, we explain the network architecture and training scheme in details for the application of multi-person pose estimation. The network for pose estimation also consists of four parts: 1) a front end CNN, built based on the architecture of Cao et al. (Cao et al., 2017) that outputs feature representations for body joints and limbs; 2) two fully connected layers to convert the features to the unary potentials of the CRF model; 3) a stack of customized layers that perform the iterative mean field updates with high-order potentials; and 4) a standard cross entropy layer to penalize the predicted edge label that is inconsistent with the ground truth label.

### A.1.1 FROM CNN TO UNARY POTENTIALS

**Front-end CNN.** Recent work He et al. (2017); Pishchulin et al. (2016); Cao et al. (2017); Insafutdinov et al. (2017) has made significant progress on multi-person pose estimation by the driving force of deep feature learning. For instance, the work proposed by Cao et al. Cao et al. (2017) presents an effective 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 pose. We utilize their network architecture as the front end CNN. Our model is complementary to this work in the way that our focus is the joint optimization of the deep feature learning and the detection association. The network proposed in Cao et al. (2017) has two separate branches after sharing several convolutional layers: one branch predicts the confidence maps for 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 Cao et al. (2017), the part fields are built 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 jump connections, 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 the joint, we keep multiple detection hypotheses even for those that are in close proximity. A detection graph is then constructed in the way that we insert edges for pairs of hypotheses that describe the same type of body joint, and for pairs of hypotheses between two different joints. Note that, although the constructed graph is not fully connected, every chordless cycle in the graph consists of only three edges.

**Edge Feature.** One of the keys to 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, the corresponding part field estimation is utilized as features. More specifically, we compute the inner product between the unit vector defined by the direction of the edge and vectors estimated by the part field. We collect a constant number of values by uniformly sampling along the line segment defined by the edge to form the feature $f_e$ for the corresponding edge. For the edges connecting the detection hypotheses of the same joint type, we simply use the euclidean distance between the detection as the feature.

**The Unary $\psi^U$.** It is straightforward to construct the unary potentials $\psi_i^U(x_i)$ (Eq. 5) from the edge feature $f_e$. We incorporate several 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. 2.3, during training, we can obtain the error signal from the mean field updates to learn the parameters of the newly introduced fully connected layers and the front end CNN that produces the edge feature.

### A.1.2 MEAN FIELD UPDATES

Zheng et al. Zheng et al. (2015) propose to formulate the mean field iteration as recurrent neural network layers, and Arnab et al. (2016) 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 our pattern-based potentials. The goal of the mean field iterations is to update the marginal distribution $Q_i^t(x_i = l)$. For initialization, $Q_i^1(x_i = l) = \frac{1}{Z_i}\exp\{-\psi_i^U(x_i = l)\}$, where $Z_i = \sum_l \exp\{-\psi_i^U, (x_i = l)\}$ is performed. 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 $\gamma_{max}$ 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 Cao et al. (2017) as an additional step to come back to the feasible set.

### A.1.3 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 Cao et al. (2017). 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.