Multi-Human Parsing With a Graph-based Generative Adversarial Model

JIANSHU LI[*], National University of Singapore
JIAN ZHAO[*], Institute of North Electronic Equipment
CONGYAN LANG, Beijing Jiaotong Univeristy
YIDONG LI, Beijing Jiaotong Univeristy
YUNCHAO WEI, University of Technology Sydney
GUODONG GUO, IDL, Baidu Research
TERENCE SIM, National University of Singapore
SHUICHENG YAN, Yitu Technology
JIASHI FENG, National University of Singapore

Human parsing is an important task in human-centric image understanding in computer vision and multimedia systems. However, most existing works on human parsing mainly tackle the single-person scenario, which deviates from real-world applications where multiple persons are present simultaneously with interaction and occlusion. To address such a challenging multi-human parsing problem, we introduce a novel multi-human parsing model named MH-Parser, which uses a graph-based generative adversarial model to address the challenges of close person interaction and occlusion in multi-human parsing. To validate the effectiveness of the new model, we collect a new dataset named Multi-Human Parsing (MHP), which contains multiple persons with intensive person interaction and entanglement. Experiments on the new MHP dataset and existing datasets demonstrate that the proposed method is effective in addressing the multi-human parsing problem compared with existing solutions in the literature.

CCS Concepts: • **Computing methodologies → Image segmentation**; *Activity recognition and understanding*; *Supervised learning*; *Multi-task learning*; *Neural networks*; Biometrics; Scene understanding; Image representations.

Additional Key Words and Phrases: Human Parsing, Multi-Human Parsing, Human-Centric Image Analysis, Generative Adversarial Networks, Graph Convolution Network

ACM Reference Format:
Jianshu Li, Jian Zhao, Congyan Lang, Yidong Li, Yunchao Wei, Guodong Guo, Terence Sim, Shuicheng Yan, and Jiashi Feng. 2020. Multi-Human Parsing With a Graph-based Generative Adversarial Model. *J. ACM* 37, 4 (August 2020), 21 pages. https://doi.org/10.1145/1122445.1122456

[*] Both authors contributed equally to this research.

Authors' addresses: Jianshu Li, jianshu@u.nus.edu, National University of Singapore, 13 Computing Drive, Singapore, 117417; Jian Zhao, zhaojian90@u.nus.edu, Institute of North Electronic Equipment; Congyan Lang, Beijing Jiaotong Univeristy, cylang@bjtu.edu.cn; Yidong Li, Beijing Jiaotong Univeristy, ydli@bjtu.edu.cn; Yunchao Wei, University of Technology Sydney, wychao1987@gmail.com; Guodong Guo, IDL, Baidu Research, guodong.guo@mail.wvu.edu; Terence Sim, National University of Singapore, tsim@comp.nus.edu.sg; Shuicheng Yan, Yitu Technology, eleyans@nus.edu.sg; Jiashi Feng, National University of Singapore, elefjia@nus.edu.sg.

0004-5411/2020/8-ART $15.00
https://doi.org/10.1145/1122445.1122456

1 INTRODUCTION

Human parsing refers to partitioning persons captured in an image into multiple semantically consistent regions, *e.g.* body parts and clothing items. As a fine-grained semantic segmentation task, it is more challenging than human segmentation which aims to find silhouettes of persons. Human parsing is very important for human-centric analysis and has lots of industrial applications, *e.g.* virtual reality, video surveillance, and human behavior analysis [17, 37]. Recently, more research attention [18, 24, 32] is paid to the more realistic scenario of multi-human parsing, where multiple persons are simultaneously present. Multi-human parsing is much more challenging than human parsing, especially when persons have close interactions in the image.

To tackle multi-human parsing, most existing methods focus on single human parsing and rely on separate off-the-shelf person detectors to localize persons in images. Single human parsing can be solved by semantic segmentation alike methods [38]. When combining the single human parsing results with the predictions from person detection or person instance segmentation methods [11, 21, 33], one can obtain the results of the multi-human parsing task. However, these methods often fail when persons have close interactions or occlusions with each other in the image, as the person detection will fail to distinguish persons in such cases. Unlike these existing methods, we propose a novel model named Multi-Human Parser (MH-Parser) to address the challenge of close person entanglement in the multi-human parsing problem. The proposed MH-Parser tackles multiple human parsing by generating global parsing maps and instance masks for multiple persons simultaneously in a bottom-up fashion, without resorting to any ad-hoc detection models. To better capture the human body structures, part configuration and human interactions, the proposed MH-Parser introduces a novel graph-based Generative Adversarial Network (Graph-GAN) model that learns to predict graph-structured instance parsing results by developing a graph-based convolutional discriminative model. Using such a graph-based discriminative model, MH-Parser is able to generate much better person instance separation results for the task of multi-human parsing.

To verify the effectiveness of our proposed MH-Parser, we collect a new dataset named Multi-Human Parsing (MHP). The person instances in MHP are entangled with close interactions and occlusion, which aligns with most realistic application scenarios. Specifically, there are 4,980 images in the dataset, and every image has 2-16 persons (3 on average) persons. For each person instance, 18 pre-defined semantic categories (also commonly used in single human parsing datasets) are annotated, including *hat*, *hair*, *sun glasses*, *upper clothes*, *skirt*, *pants*, *dress*, *belt*, *left shoe*, *right shoe*, *face*, *left leg*, *right leg*, *left arm*, *right arm*, *bag*, *scarf* and *torso skin*. Each instance has a complete set of annotations whenever the corresponding category is present in the image. Thus MHP is a much more challenging dataset and will serve as a more realistic benchmark on human-centric analysis to push the frontier of human parsing research.

To sum up, we make the following contributions. 1) We propose a novel model MH-Parser, which uses a graph-based discriminative model to address the close person entanglement problem in multi-human parsing. 2) We construct the MHP dataset, a new multi-human parsing benchmark, which has more person entanglement and matches real-world scenarios better. 3) We conduct extensive experiments on the multi-human parsing task to push the frontier of multi-human parsing techniques.

2 RELATED WORK

2.1 Human Parsing

Previous human parsing methods [36, 38, 41] and datasets [9, 35, 38, 55] mainly focus on single-human parsing, which have severe practical limitations. None of the commonly used human parsing datasets considers instance-aware cases. Moreover, the persons in these datasets are usually in upright positions with limited pose changes, which does not accord with reality. Recently, human parsing in the wild is inspected in [18], where persons present varying clothing appearances and diverse viewpoints, but it only considers the setting of instance-agnostic human parsing. Different from existing datasets on human parsing, the proposed MHP dataset considers simultaneous presence of multiple persons in an instance-aware setting with challenging pose variations, occlusion and interaction between persons, aligning much better with reality.

2.2 Instance-Aware Object/Human Segmentation

Recently, many research efforts have been devoted to instance-aware object/human semantic segmentation. It can be solved by top-down approaches and bottom-up approaches. In the top-down family, a detector (or a component functioning as a detector) is used to localize each instance, which is further processed to generate pixel segmentation. Multi-task Network Cascades (MNC) [11] consists of three separate networks for differentiating instances, estimating masks and categorizing objects, receptively. The first fully convolutional end-to-end solution to instance-aware semantic segmentation in [33] performs instance mask prediction and classification jointly. Mask-RCNN [21] adds a segmentation branch to the state-of-the-art object detector Faster-RCNN [48] to perform instance segmentation. The top-down approaches heavily depend on the detection component, and suffer poor performance when instances are close to each other. In the bottom-up family, detection is usually not used. Usually embeddings of all pixels are learned, which are later used to cluster different pixels into different instances. In [45], embeddings are learned with a grouping loss, which does pairwise comparisons across randomly sampled pixels. In [12, 44], a discriminative loss containing push forces and pull forces is used to learn embeddings for each pixel. In [37], the embeddings of pixels are learned with direct supervision of instance locations. Different from the methods which operate on pixels, we learn an embedding of each superpixel. Furthermore, Graph-GAN is used to refine the learned embedding by leveraging high-order information. These instance-aware person segmentation methods, either top-down or bottom-up approaches, can only predict person-level segmentation without any detailed information on body parts and fashion categories, which is disadvantageous for fine-grained image understanding. In contrast, our MHP is proposed for fine-grained multi-human parsing in the wild, which aims to boost the research in real-world human-centric analysis.

2.3 Generative Adversarial Networks

The recently proposed GAN-based methods [3, 19, 47] have yielded remarkable performance on generating photo-realistic images [22] and semantic segmentation maps [42] by specifying only a high-level goal like “to make the output indistinguishable from the reality” [23]. GAN automatically learns a customized loss function that adapts to data and guides the generation process of high-quality images. Different from existing works on image-based GANs which can only process regular input (*e.g.* 2D grid images), the discriminator in our model takes a flexible data structure, *i.e.* graphs, as input, and aims to improve the realism of the input graph-structured data. This is made possible by the recent advances of graph-based algorithms [53] and graph convolution networks [26]. Using graph networks in a GAN setting has been explored in [6, 27, 54]. However, these works are devoted for tasks like graph generation, link prediction and node classification or embedding. Our work is the first one to explore how to use graph-based GAN in the area of pixel-level human-centric image understanding.

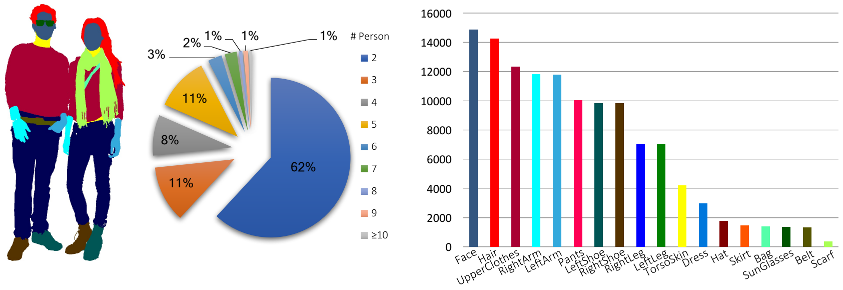

Fig. 1. Examples and statistics of the MHP dataset. Left: An annotated example for multi-human parsing. Middle: Statistics on number of persons in one image. Right: The data distribution on 18 semantic part labels in the MHP dataset.

3 THE MHP DATASET

In this section we introduce the Multiple Human Parsing (MHP) dataset designed for multi-human parsing in the wild.

3.1 Image Collection and Annotation Methodology

As pointed out in [24], in generic recognition datasets like PASCAL [14] or COCO [39], only a small percentage of images contain multiple persons. Also, persons in these generic recognition datasets usually lack fine details, compared to human-centric datasets, such as those for people recognition in photo album [46], human immediacy prediction [10], interpersonal relation prediction [56], *etc.* To benefit the development of new multi-human parsing models, we construct a pool of images from existing human-centric datasets [10, 46, 51, 56], and also online Creative Commons licensed imagery. From the images pool, we select a subset of images which contain clearly visible persons with intimate interaction, rich fashion items and diverse appearances, and manually annotate them with two operations: 1) counting and indexing the persons in the images and 2) annotating each person. We implement an annotation tool and generate multi-scale superpixels of images based on [2] to speed up the annotation. For each instance, 18 pre-defined semantic categories (also commonly used in single-parsing datasets) are annotated, including *hat*, *hair*, *sun glasses*, *upper clothes*, *skirt*, *pants*, *dress*, *belt*, *left shoe*, *right shoe*, *face*, *left leg*, *right leg*, *left arm*, *right arm*, *bag*, *scarf* and *torso skin*. Each instance has a complete set of annotations whenever the corresponding category is present in the image. When annotating one instance, others are regarded as background. Thus, the resulting annotation set for each image consists of P person-level parsing masks, where P is the number of persons in the image.

3.2 Dataset Statistics

MHP dataset contains various numbers of persons in each image, and the distribution is illustrated in Fig. 1 (middle). Real-world human parsing aims to analyze every detailed region of each person of interest, including different body parts, clothes and accessories. Thus we define 7 body parts and 11 clothing and accessory categories. Among these 7 body parts, we divide *arms* and *legs* into left and right side for more precise analysis, which also increases the difficulty of the task. As for clothing categories, we have not only common clothes like *upper clothes*, *pants*, and *shoes*, but also confusing categories such as *skirt* and *dress* and infrequent categories such as *scarf*, *sun glasses*, *belt*, and *bag*. The statistics for each semantic part annotation are shown in Fig. 1 (right).

In the MHP dataset, there are 4,980 images, each with multiple persons, each with 2-16 persons (3 on average). The resolution of the images ranges from 284 × 117 to 6,919 × 4,511, with an average

Table 1. Comparison with publicly available datasets for human parsing. For each dataset we report the total number of images, the number of categories including background, the average number of persons per image and the extent of occlusion between persons, quantified by the average IOU between person pairs.

Dataset	No. of Images	No. of Categories	No. of Persons	Extent of Occlusion
Fashionista [55]	685	56	1	N.A.
ATR [38]	17,700	18	1	N.A.
LIP [18]	50,462	20	1	N.A.
PASCAL-Person-Part [8, 9]	3,533	7	2.19	4.60%
Buffy [52]	748	13	2.36	4.17%
MHP (ours)	4,980	19	3.01	11.71%

of 755 × 734 pixels. Totally there are 14,969 person instances with fine-grained annotations at pixel-level with 18 different semantic labels. The resolution of each person ranges from 64 × 43 to 2,627 × 3,881, with an average 224 × 565 pixels. For other human parsing datasets, Fashionista [55] contains 685 person instances, ATR [38] contains 17,700 and LIP [18] contains 50,462. However, they all reflect the cases of single-human parsing, which deviates from real-world human parsing requirement. Originally the PASCAL-Person-Part dataset is proposed in [8, 9] for semantic part segmentation. It was first used in an instance-level part segmentation setting in [32]. Although this dataset only has coarse human part categories while our MHP dataset has fine-grained categories designed for fashion analysis, they share common philosophy of parsing person into parts while differentiating person instances. Thus we also compare with this PASCAL-Person-Part dataset here.

In MHP, the person instances are entangled with close interaction and occlusion. To verify this, we calculate the mean average Intersection Over Union (IOU) of person bounding boxes in the dataset. That is, we find the average IOU between person instances in each image, and calculate its mean value over the whole dataset. In MHP the mean average IOU is 11.71%. As a widely used human instance segmentation dataset, COCO [39] only has mean average IOU of 2.81% for the images with multiple persons. Even for the Buffy [52] dataset, which is used in person individuation and claims to have multiple closely entangled persons [24], the mean average IOU is only 4.17%. The details of the comparisons are summarized in Tab. 1. Thus MHP is a much more challenging dataset in terms of separating closely entangled person instances. Therefore, the MHP dataset will serve as a more realistic benchmark on human-centric analysis to push the frontier of human parsing research.

4 THE MH-PARSER

In this section we elaborate on the proposed MH-Parser model for parsing multiple persons. The proposed MH-Parser simultaneously generates a global semantic parsing map and a pairwise affinity map (which is used to construct instance masks). The former presents union of the instance parsing maps for all the persons in the input image, and the latter distinguishes one person from another. The overall architecture of MH-Parser is shown in Fig. 2.

4.1 Global Parsing Prediction

The MH-Parser uses a deep representation learner to learn rich and discriminative representations which are shareable for global parsing and affinity map prediction. In particular, the representation learner is a fully convolutional network consisting of 101 layers (ResNet101) adopted from DeepLab [8]. It generates features with 1/8 of the spatial dimension of the input image. On top of

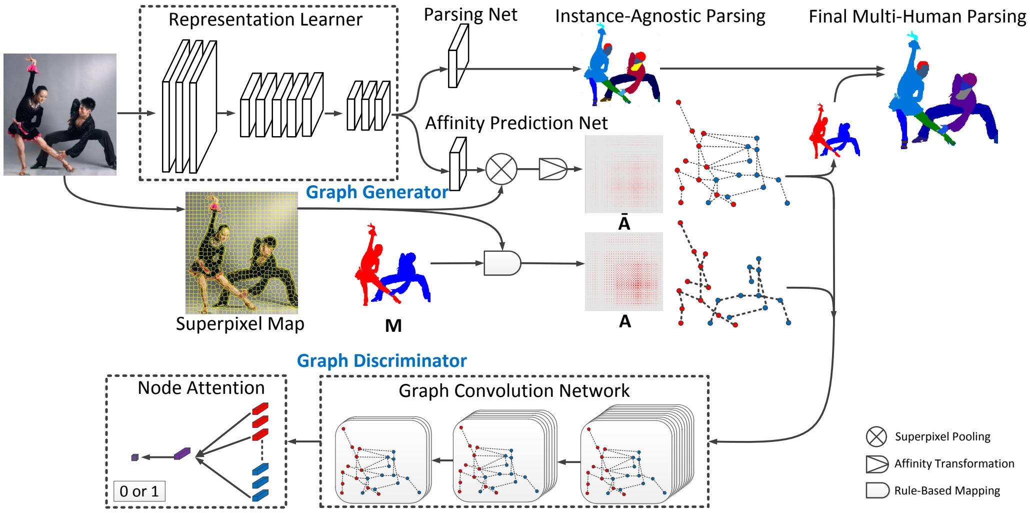

Fig. 2. Architecture overview of the proposed Multiple Human Parser (MH-Parser). Here $\mathbf{M}$ refers to the global accordance map, $\mathbf{A}$ refers to the ground truth pairwise affinity map and $\bar{\mathbf{A}}$ denotes the predictions. $\mathbf{A}$ is obtained by rule-based mapping from $\mathbf{M}$ and the corresponding superpixel map (see Eqn. (4) and (5)), and $\bar{\mathbf{A}}$ is the output of the graph generator (consisting of the representation learner and the affinity prediction net). The graph convolution discriminator takes the affinity graph from the graph generator as input and predicts whether it is a ground truth or a prediction. Fusing the predicted instance-agnostic parsing map and instance masks (constructed from $\bar{\mathbf{A}}$) gives the instance-aware parsing results.

this learner, a small parsing net consisting of atrous spatial pyramid pooling [8] is used to generate instance-agnostic semantic parsing maps of the whole image, as shown in Fig. 2.

Formally, let G_{seg} denote the global parsing module. Given an input image $\mathbf{I}$ with size $H \times W$, its output $\bar{\mathbf{S}} = G_{\text{seg}}(\mathbf{I}) \in \mathbb{R}^{H' \times W' \times C}$ gives instance-agnostic parsing of C categories with a scaled down size compared with the input image $\mathbf{I}$. The global parsing predictor G_{seg} can be trained by minimizing the following standard parsing loss:

$$\mathcal{L}_{\text{seg}}(G) \triangleq \mathcal{L}_{\text{ce}}(G_{\text{seg}}(\mathbf{I}), \mathbf{S}), \tag{1}$$

where $\mathcal{L}_{\text{ce}}$ is the pixel-wise cross-entropy loss and $\mathbf{S}$ is the ground truth labeling of the instance-agnostic semantic parsing map.

4.2 Graph-GAN for Affinity Map Prediction

The global parsing results do not present any instance-level information which however is essential for multi-human parsing. Different from top-down solutions, we propose a novel graph-GAN model for learning instance information in a bottom-up fashion simultaneously with the global parsing prediction.

4.2.1 Global Accordance Map. Global accordance maps distinguish different persons by associating them with different accordance scores. For an input image $\mathbf{I}$ with size $H \times W$, its global accordance map $\mathbf{M} \in \mathbb{R}^{H \times W}$ is defined as

$$\mathbf{M}(k) = \begin{cases} i, & \text{if pixel } k \text{ is from the } i\text{-th person,} \\ 0, & \text{otherwise.} \end{cases} \tag{2}$$

An example of the global accordance map **M** constructed from the ground truth instance parsing map is shown in Fig. 2.

Predicting global accordance scores accurately is important for separating different person instances and deriving high-quality multi-human parsing results. However, accordance prediction is very challenging, due to the large appearance variance of intra-instance pixels and subtle difference of some pixels from different instances. The number of persons is unknown and varies for different images, making traditional classification approaches inapplicable. Moreover, the accordance scores are expected to be invariant to permutation over person instance ids. This implies that the learning process of accordance score is extremely unstable if we directly use the ground truth global accordance map defined in Eqn. (2) as supervision.

4.2.2 Pairwise Affinity Graph. Since directly predicting global accordance scores is difficult, MH-Parser generates a pairwise affinity graph instead. Under the context of multi-human pose estimation, such pairwise affinity between person key-points are used in [7]. Here in multi-human parsing, the MH-Parser introduces a graph generator to learn to optimize the pairwise distances (or affinities) among regions within input images. In MH-Parser, superpixel is regarded as the basic unit of regions to calculate the affinities, due to the following two reasons. First, superpixels are natural low-level representations to delineate boundaries between semantic concepts. Second, superpixels can be regarded as low-level pixel grouping, so that the complexity of affinity computation is greatly reduced compared to pixel level affinity computation.

Formally, we define the pairwise affinity graph as

$$\mathcal{G} = (\mathcal{V}, \mathcal{E}), \quad \begin{cases} v_n = s_n, \forall n \in [1, 2 \cdots N], \\ e_{n_1,n_2} = \mathbf{A}(n_1, n_2). \end{cases} \tag{3}$$

In the graph, each vertex $v_n \in \mathcal{V}$ is one superpixel within the image. There are N superpixels in total and s_n is the n-th superpixel. Each edge $e_{n_1,n_2} \in \mathcal{E}$ represents the connectivities between each pair of vertices (v_{n_1}, v_{n_2}), described by the pairwise affinity map **A**. The ground truth pairwise affinity map $\mathbf{A} \in \mathbb{R}^{N\times N}$ is derived from a rule-based mapping, which is defined as

$$\mathbf{A}(n_1, n_2) = \begin{cases} 1, \text{if } \sigma_{\text{gt}}(s_{n_1}) = \sigma_{\text{gt}}(s_{n_2}) \text{ and } \sigma_{\text{gt}}(s_{n_1}) > 0, \\ 0, \text{otherwise}, \end{cases} \tag{4}$$

and

$$\sigma_{\text{gt}}(s_n) = \bigwedge_{k\in s_n} \mathbf{M}(k). \tag{5}$$

Here $k \in s_n$ represents all pixels within s_n and $\bigwedge$ denotes the majority vote operation. Note that although the ground truth **M** has multiple possible values due to random assignment of person ids, the corresponding ground truth **A** is unique regardless of how the person ids are assigned.

The pairwise affinity maps can be learned directly by taking the ground truth **A** as the regression target. The predicted pairwise affinity map $\bar{\mathbf{A}}$ can be generated directly by an affinity prediction net, which draws features from the representation learner. The affinity prediction net first generates a set of features $\mathbf{F} \in \mathbb{R}^{H''\times W''\times C_F}$, where C_F is the number of channels for **F**. Then it applies superpixel pooling on **F**, followed by an affinity transformation with a Gaussian kernel to obtain $\bar{\mathbf{A}} \in \mathbb{R}^{N\times N}$:

$$\bar{\mathbf{A}}(n_1, n_2) = \exp\left(-\frac{\sum_{c=1}^{C_F} \left[\sigma_{\text{sp}}(s_{n_1}, c) - \sigma_{\text{sp}}(s_{n_2},c)\right]^2}{2\theta^2} \right), \tag{6}$$

where

$$\sigma_{\mathrm{sp}}(s_n, c) = \frac{1}{\|s_n\|} \sum_{k \in s_n} \mathbf{F}(k, c). \tag{7}$$

Here θ is the parameter controlling sensitivity of $\bar{\mathbf{A}}$, and $\|s_n\|$ is the number of pixels within superpixel s_n.

The network for predicting $\bar{\mathbf{A}}$ can be trained by minimizing the distance between $\bar{\mathbf{A}}$ and its ground truth $\mathbf{A}$. However, when learning $\bar{\mathbf{A}}$ with direct supervision, the elements within it are learned independently of each other. The contiguity and relations (reflecting intrinsic human body structures) within $\bar{\mathbf{A}}$ are not captured. For example, if node v_{n_1} is connected to v_{n_2} and v_{n_2} is connected to v_{n_3}, then v_{n_1} is also connected to v_{n_3}. This higher-order affinity between regions is not captured for the case of direct supervision.

4.2.3 Predicting Affinity Graph with Graph-GAN. To remedy the potential issues in learning with direct supervision over $\mathbf{A}$, we propose a novel GAN model, Graph-GAN, to augment the learning process. Traditionally, high-order consistencies between labels are modelled by Conditional Random Field (CRF) [30] in both traditional methods [28, 50] and deep learning[4, 16]. Different from these existing works, we use Graph-GAN in our proposed method to capture such high-order information. Also, different from existing GAN-based models which can only process regular input (like 2D grid images), the Graph-GAN can take in and process flexible graph-structured data. It aims to learn high-quality affinity graphs to better capture the human body structure, part configuration and human interaction.

In the adversarial learning of the Graph-GAN model, the ground truth affinity graphs use $\mathbf{A}$ from Eqn. (4) in the edge definition. The predicted affinity graphs use $\bar{\mathbf{A}}$ from Eqn. (12). The generator in Graph-GAN learns to generate high-quality affinity graphs, which are indistinguishable from the ground truth. The discriminator in Graph-GAN targets at telling the predicted affinity graphs apart from ground truth ones. With generator and discriminator playing against each other, the discriminator learns to supervise the generator in a way tailored for the graph-structured data.

The representation learner and the affinity prediction net are adopted as the generator in the Graph-GAN model to generate the predicted affinity graph. In order to handle graph-structured input, we propose a Graph Convolution Network (GCN) based discriminator model. The GCN [13, 26, 43] can effectively model graph-structured data, thus is suitable for classifying input graphs and serves as the discriminator.

In particular, we use a simple form of layer-wise propagation rule [26, 43]:

$$\mathbf{H}^{(l+1)} = f(\mathbf{H}^{(l)}, \mathrm{A}) = \sigma(\mathbf{A}\mathbf{H}^{(l)}\mathbf{W}^{(l)} + \mathbf{b}^{(l)}), \tag{8}$$

where $\mathbf{A}$ is the adjacency matrix of the graph (pairwise affinity map in our case), $\mathbf{H}$ denotes the hidden activations in GCN, $\mathbf{W}$ and $\mathbf{b}$ denote the learnable weights and biases, σ is a non-linear activation function, and l is the layer index. Thus $\mathbf{H}^{(0)}$ represents the input node features and $\mathbf{H}^{(L)}$ represents the output node features, where L is the total number of layers in GCN. Note that we use a fixed number of superpixels, thus the dimension of $\mathbf{H}^{(0)}$ is fixed. Here the input feature to the GCN model is a one-hot embedding of each node, *i.e.* $\mathbf{H}^{(0)} \in \mathbb{R}^{N \times N} = \mathbf{I}_N$. We follow [26] and normalize the adjacency matrix to make the propagation stable:

$$\mathbf{H}^{(l+1)} = \sigma(\hat{\mathbf{D}}^{-\frac{1}{2}}\hat{\mathbf{A}}\hat{\mathbf{D}}^{-\frac{1}{2}}\mathbf{H}^{(l)}\mathbf{W}^{(l)} + \mathbf{b}^{(l)}), \tag{9}$$

where $\hat{\mathbf{A}} = \mathbf{A} + \mathbf{I}_N$ with $\mathbf{I}_N$ as the identity matrix and $\hat{\mathbf{D}}$ is the diagonal node degree matrix of $\hat{\mathbf{A}}$, *i.e.* $\hat{\mathbf{D}}_{ii} = \sum_j \hat{\mathbf{A}}_{ij}$.

In GCN, the graph convolution operation effectively diffuses the features across different regions (including body parts and background) based on the connectivities between the regions. With multiple layers of feature propagation within GCN, higher order relations of different regions are captured, which help identify the intrinsic body part structures of multiple humans.

Since the layer propagation rule in Eqn. (9) only models the transformation of the features of nodes, node pooling operation is defined in order to obtain a graph-level feature. We define a node pooling layer on top of the final output node features with an attention mechanism, as usually used in nature language processing [34, 40]:

$$\mathbf{H}_g = \sum_{n=1}^{N} w_n^{\mathrm{att}} \mathbf{H}_n^{(L)}, \tag{10}$$

where

$$w^{\mathrm{att}} = \mathrm{softmax}(\mathrm{atten}(\mathbf{H}^{(\mathrm{att})})) \in \mathbb{R}^{\mathrm{N}}. \tag{11}$$

Here w^{att} is the attention weight vector, which is generated by a simple linear layer when taking the feature $\mathbf{H}^{(\mathrm{att})}$ as input. In this work we use $\mathbf{H}^{(\mathrm{att})} \in \mathbb{R}^{N \times C}$ as the input to the node attention layer, which is defined as

$$\mathbf{H}^{\mathrm{att}}(n, c) = \frac{1}{\|s_n\|} \sum_{k \in s_n} \bar{\mathrm{S}}(k, c), \text{ for } c \in \{1, 2, \cdots C\}. \tag{12}$$

Specifically, $\mathbf{H}^{(\mathrm{att})}$ is the feature from the parsing net prediction with superpixel pooling operations applied to it, such that each node corresponds to a C-dimensional feature vector.

Essentially, the node attention operation multiplies w_n^{att} (the n-th element of the attention weight vector w^{att}) with its corresponding node feature vector $\mathbf{H}_n^{(L)}$ (the n-th node's from the last layer of the graph), and then sums up the weighted node feature vectors. We use the attention mechanism as the node feature pooling, resulting in a single descriptor $\mathbf{H}_g$ for the whole graph. With the attention pooling operation, the diffused features of different regions within an image are aggregated into one feature vector. Then $\mathbf{H}_g$ is used as the input to a classifier to predict whether the input affinity graph is a ground truth one or a predicted one in the adversarial training setting.

4.3 Training and Inference

We train the generator by introducing the following losses. For the global parsing task, the loss function in Eqn. (1) is used. For the affinity graph prediction task, we minimize the distance between the predicted pairwise affinity map $\bar{\mathbf{A}}$ and the ground truth pairwise affinity map $\mathbf{A}$ with $L2$ loss:

$$\mathcal{L}_{L2}(G) = \|\mathbf{A} - G_{\mathrm{graph}}(\mathbf{I}) \odot \mathbf{A}_{\mathrm{fg}}\|^2. \tag{13}$$

Here $G_{\mathrm{graph}}(\cdot)$ represents the mapping function from the input image $\mathbf{I}$ to the predicted pairwise affinity map, *i.e.* $\bar{\mathbf{A}} = G_{\mathrm{graph}}(\mathbf{I})$, and $\odot$ denotes the element-wise multiplication operator. This mapping function is realized by the representation learner and affinity prediction net in Fig. 2. The foreground mask $\mathbf{A}_{\mathrm{fg}}$ is a binary mask indicating connections only between foreground nodes. It is used to set all other connections to 0 and is defined as

$$\mathbf{A}_{\mathrm{fg}}(n_1, n_2) = \begin{cases} 1, \text{if } \sigma_{\mathrm{gt}}(s_{n_1}) > 0 \text{ and } \sigma_{\mathrm{gt}}(s_{n_2}) > 0, \\ 0, \text{otherwise.} \end{cases} \tag{14}$$

For training the Graph-GAN, the corresponding loss is

$$\mathcal{L}_{\mathrm{GAN}}(G, D) = \log(D(\mathbf{A})) + \log(1 - D(G_{\mathrm{graph}}(\mathbf{I}) \odot \mathbf{A}_{\mathrm{fg}})), \tag{15}$$

where D denotes the GCN-based discriminator. Thus the overall objective function is to find G^* such that

$$G^* = \arg\min_G \max_D \mathcal{L}_{\mathrm{seg}}(G) + \mathcal{L}_{L2}(G) + \lambda \mathcal{L}_{\mathrm{GAN}}(G, D). \tag{16}$$

After finding the optimal G^*, we use it to generate global parsing maps and affinity maps for testing images.

During testing, we use the predicted affinity graph $\bar{\mathbf{A}}$ to perform spectral clustering. Background nodes are identified by the global parsing map, and are removed from the affinity graph. Then all the foreground nodes are clustered according to the pairwise affinities in $\bar{\mathbf{A}}$. Different instances of persons are identified from the clustering results. To help clustering, a regression layer built upon the representation learner is used to learn the number of persons during training, and the predicted person number is used in clustering during testing. It is omitted in the network structure and the objective function for brevity.

4.4 Instance Mask Refinement

We extend our model with a refinement step to reinforce the prediction of instance masks (obtained from the clustering results) from superpixel level to pixel level. We adopt CRF [30] in refinement to associate each pixel in the image with one of the persons (from the clustering results) or background. The CRF model contains two unary terms, *i.e.* $\Psi_u = \Psi_{\text{Person}} + \Psi_{\text{Global}}$ and a binary term Ψ_p. With V_k denoting the random variable for the k-th pixel in the image, the target of the instance mask refinement is to find the optimal solution V_k for all pixels in the image that minimizes the following energy function:

$$E = -\sum_k \ln \Psi_u(V_k) + \sum_{k_1<k_2} \Psi_p(V_{k_1}, V_{k_2}). \tag{17}$$

We define these terms as follows. Given P persons from clustering results over the predicted affinity map, and assuming the i-th person is represented by a binary mask $\mathbf{P}^i$ indicating whether the pixel is from the i-th person, we define the person consistency term Ψ_{Person} as

$$\Psi_{\text{Person}}(V_k = i) = \mathbf{Q}_k \mathbf{P}^i_k, \text{ for } V_k \in [0, 1, 2, \cdots P]. \tag{18}$$

Here $\mathbf{Q}_k$ denotes the probability of the k-th pixel to be foreground. The person consistency term is designed to give strong cues about which person each foreground pixel should belong to. As in [5, 32], the global term Ψ_{Global} is defined as

$$\Psi_{\text{Global}}(V_k = i) = \mathbf{Q}_k, \tag{19}$$

which is used to complete the person consistency term by giving equal likelihood of each foreground pixel to all the persons to correct errors in the clustering process. Finally we define our pairwise term as

$$\Psi_p(V_{k_1}, V_{k_2}) = \mu(V_{k_1}, V_{k_2})\kappa(\mathbf{f}_{k_1}, \mathbf{f}_{k_2}), \tag{20}$$

where $\mu(\cdot, \cdot)$ is the compatibility function, $\kappa(\cdot, \cdot)$ is the kernel function and $\mathbf{f}_k$ is the feature vector at spatial location k. The feature vector contains the C_F-dimensional vector from $\hat{\mathbf{F}}(k)$ (obtained by up-sampling $\mathbf{F}$ to match the spatial dimension of the input image) in the affinity prediction net, the 3-dimensional color vector $\mathbf{I}_k$, and the 2-dimensional position vector $\mathbf{p}_k$. Thus the kernel is defined as

$$\begin{aligned}\kappa(\mathbf{f}_{k_1}, \mathbf{f}_{k_2}) = w^{(1)} &\exp\left(-\frac{\|\hat{\mathbf{F}}(k_1) - \hat{\mathbf{F}}(k_2)\|^2}{2\theta^2}\right) \\ &+ w^{(2)} \exp\left(-\frac{\|\mathbf{p}_{k_1} - \mathbf{p}_{k_2}\|^2}{2\theta^2_{b_p}} - \frac{\|\mathbf{I}_{k_1} - \mathbf{I}_{k_2}\|^2}{2\theta^2_{b_I}}\right) \\ &+ w^{(3)} \exp\left(-\frac{\|\mathbf{p}_{k_1} - \mathbf{p}_{k_2}\|^2}{2\theta^2_s}\right).\end{aligned} \tag{21}$$

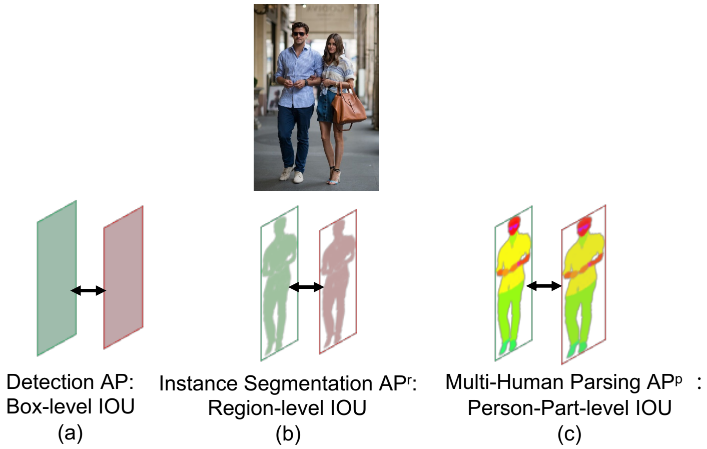

Fig. 3. The illustration of different types of IOU used in AP, AP^r and AP^p. (a) AP for object detection uses box-level IOU; (b) AP^r for instance segmentation uses region-level IOU; (c) AP^p for multi-human parsing uses person-part-level IOU.

In other words, the pairwise kernel consists of the learned features for pairwise distance measurement, in addition to the bilateral term and the spatial term used in [29]. The compatibility function is realized by the simple Potts model.

With the above CRF model, we find the optimal solution that minimizes the energy function in Eqn. (17) with the approximation algorithm in [29] and obtain the final prediction of person instance masks for each pixel in input images. Standard CRF is also applied to the instance-agnostic parsing maps as in [8].

5 EXPERIMENTS

5.1 Experimental Setup

5.1.1 Performance Evaluation Metrics. We use the following performance evaluation metrics for multi-human parsing.

Average Precision based on Part (AP^p). Different from region-based Average Precision (AP^r) used in instance segmentation [20, 37], AP^p uses part-level Intersection Over Union (IOU) of different semantic part categories within a person to determine if one instance is a true positive. Specifically, when comparing one predicted semantic part parsing map with one ground truth parsing map, we find the IOU of all the semantic part categories between them and use the average as the measure of overlap. We refer to AP under this condition as AP^p. We prefer AP^p over AP^r, as we focus on human-centric evaluation and we pay attention to how well a person as a whole is parsed. The comparison of the IOU used in AP (for object detection), AP^r (for instance segmentation) and AP^p (for multi-human parsing) is shown in Fig 3. Moreover, similarly to AP$^r_{vol}$, we use AP$^p_{vol}$ to denote the average AP^p values at IOU threshold from 0.1 to 0.9 with a step size of 0.1.

Percentage of Correctly Parsed Body Parts (PCP). As AP^p averages the IOU of each part category, it cannot reflect how many parts are correctly predicted. Thus we propose to adopt PCP, originally used in human pose estimation [9, 15], to evaluate parsing quality on the semantic parts within person instances. For each true-positive person instance, we find all the categories (excluding background) with pixel-level IOU larger than a threshold, which are regarded as correctly parsed. PCP of one person is the ratio between the correctly parsed categories and the total number of categories of that person. Missed person instances are assigned 0 PCP. The overall PCP is the average PCP for all person instances. Note that PCP is also a human-centric evaluation metric.

5.1.2 Datasets. We perform experiments on the MHP dataset. From all the images in MHP, we randomly choose 980 images to form the testing set. The rest form a training set of 3,000 images and a validation set of 1,000 images. Since we are interested in the real-world situation where different people are near to each other with close interaction, we also perform experiments on the Buffy [52] dataset as suggested in [24], which contains entangled people in almost all testing images. Experiments on PASCAL-Person-Part dataset are also performed to compare with the results in [32] to validate the effectiveness of our proposed method.

5.2 Implementation Details

5.2.1 Superpixel Map. To extract superpixels, we first resize the shorter side of input images to 600 pixels and ensure that its longer side is not larger than 1,000 pixels. For each resized input image, we over-segment it into $N = 1,000$ superpixels using SLIC [1]. If the number of superpixels is not equal to 1,000, we either split or merge the last few superpixels to produce exactly 1,000 superpixels. With such a superpixel map and the ground truth person instance map, we generate the ground truth pairwise affinity map $\mathbf{A}$ and the foreground mask $\mathbf{A}_{\text{fg}}$ and cache them to accelerate the training process.

5.2.2 Network Architecture.

Input Images. For the MHP dataset, we resize the images to the same size when extracting superpixel maps as the input to the network. For images in Buffy, original size is used without the resizing operation. No data augmentation is used.

Graph Generator. The representation learner in the graph generator is a fully convolutional network consisting of 101 layers adopted from DeepLab [8]. The representation learner generates feature maps with 1/8 the spatial dimension of the input image. The parsing net generates parsing prediction of the same size. For the affinity prediction net, we use one convolutional layer followed by one bilinear up-sample layer to generate feature maps with 1/4 the spatial dimension of the input image to perform superpixel pooling, and the number of channels is set as 4.

Graph Discriminator. In the experiments, the graph discriminator uses featureless nodes with embedding. The input feature to the graph $\mathbf{H}^{(0)} \in \mathbb{R}^{N \times N}$ takes $\mathbf{I}_N$, and the feature of each node is a one-hot embedding. The graph convolution also has 5 layers, with feature dimension 1,000, 512, 512, 256 and 256, respectively. Namely, $\mathbf{H}^{(1)} \in \mathbb{R}^{N \times 1000}$, $\mathbf{H}^{(2)} \in \mathbb{R}^{N \times 512}$, $\mathbf{H}^{(3)} \in \mathbb{R}^{N \times 512}$, $\mathbf{H}^{(4)} \in \mathbb{R}^{N \times 256}$ and $\mathbf{H}^{(5)} \in \mathbb{R}^{N \times 256}$. The non-linear activations in the graph convolution is leaky-RELU with a slope of 0.3.

5.2.3 Optimization. For all the experiments, we use the pre-trained semantic segmentation model from [8] for initialization. Then the model is trained on the training set of MHP dataset. We find that pre-training both the generator and discriminator is beneficial to training the Graph-GAN model. The generator, including the representation learner, the parsing net and the affinity prediction net, is pre-trained with $\lambda = 0$, learning rate lr=10^{-4} and weight decay=5×10^{-4} for 20 epochs with polynomial learning rate policy. Once done, the parameters for the generator are fixed and the discriminator GCN is trained for another 10 epochs with lr=10^{-3}. Finally, the model is trained with $\lambda = 0$, $\theta = 0.1$ and lr=10^{-6} for the baseline $L2$ model, and with $\lambda = 0.001$, $\theta = 0.1$ and lr=10^{-6} for the Graph-GAN model. Both the generator and the discriminator are optimized with one image per batch and the optimizer used is Adam [25] with $\beta_1 = 0.5$.

Table 2. Results from different methods on the MHP test set. The results of Mask RCNN and DL are obtained by using them to predict the instance masks, respectively, and combining with the same instance agnostic parsing map produced by MH-Parser for fair comparison. All denotes the entire test set, and Top 20% and Top 5% denote two subsets of testing images with top 20% and top 5% largest overlaps between person instances, respectively.

	All			Top 20%			Top 5%		
	$\mathrm{AP}^p_{0.5}$	AP^p_{vol}	$\mathrm{PCP}_{0.5}$	$\mathrm{AP}^p_{0.5}$	AP^p_{vol}	$\mathrm{PCP}_{0.5}$	$\mathrm{AP}^p_{0.5}$	AP^p_{vol}	$\mathrm{PCP}_{0.5}$
Detect+Parse	29.81	38.83	43.78	12.08	30.22	25.44	9.76	30.37	18.36
Mask RCNN [21]	**52.68**	**49.81**	**51.87**	31.49	40.16	37.31	24.25	35.63	28.77
DL [12]	47.76	47.73	49.21	34.81	44.06	40.59	29.52	43.52	33.70
MH-Parser	50.10	48.96	50.70	**41.67**	**46.70**	**44.74**	**33.69**	**46.57**	**37.01**

For the instance mask refinement, we adopt the parameters of CRF from [8], *i.e.* $w^{(2)} = 4$, $w^{(3)} = 3$, $\theta_{b_p} = 20$, $\theta_{b_I} = 3$, $\theta_s = 3$. For the pairwise term from the affinity map, we keep $\theta = 0.1$ and set $w^{(1)}$ to 2 according to the validation set. The number of iterations of CRF optimization is 10. For CRF of refining global parsing masks, we directly use parameters from [8].

5.3 Experimental Analysis

5.3.1 Comparison with State-of-the-Arts. Note that standard instance segmentation methods can only generate silhouettes of person instances and cannot produce person part parsing as desired. Thus we use them to generate instance masks as the graph generator in MH-Parser does, and combine the instance masks with the instance agnostic parsing to produce final multi-human parsing results. Here we use Mask-RCNN [21], which is the state-of-the-art top-down model, and Discriminative Loss (DL) [12], a well established bottom-up model, to generate instance masks. For Mask RCNN, we use the segmentation prediction in each detection with high confidence (0.9) to form the instance masks. DL can generate instance masks as the outputs of the model. We also consider the Detect+Parse baseline method as used in traditional single human parsing, where a person detector is used to detect person instances, and a parser is used to parse each detected instance.

The performance of these methods in terms of AP^p, AP^p_{vol} and PCP on the MHP test set is listed in Tab. 2. In the table the overlap thresholds for AP^p and PCP are both set as 0.5. The MH-Parser, DL, the parser in Detect+Parse are trained on MHP training set with the same trunk network (ResNet101). Especially, DL is trained with the official code [12, 44] with the suggested setting. The Mask-RCNN model and the detector in Detect+Parse are the top performing model with ResNet101 as the trunk from the official Detectron [49].

We can see that the proposed MH-Parser achieves competitive performance with Mask RCNN and DL on the MHP dataset, and outperforms Detect+Parse baseline. To investigate how these models address the concerned challenges of closely entangled persons, we select two challenging subsets from the MHP test set. For each image in the test set, we perform a pairwise comparison of all the person instances, and find the IOU of person bounding boxes in each pair. Then the average IOU of all the pairs is used to measure the closeness of the persons in each image. One subset contains the images with top 20% highest average IOUs, and the other subset contains top 5%. They represent images with very close interaction of human instances, reflecting the real scenarios. The results on these two subsets are listed in Tab. 2. We can see that on these challenging subsets, MH-Parse outperforms both Mask RCNN and DL. For Mask RCNN, it has difficulties to differentiate entangled persons, while as a bottom-up approach, MH-Parse can handle such cases well. For DL, it

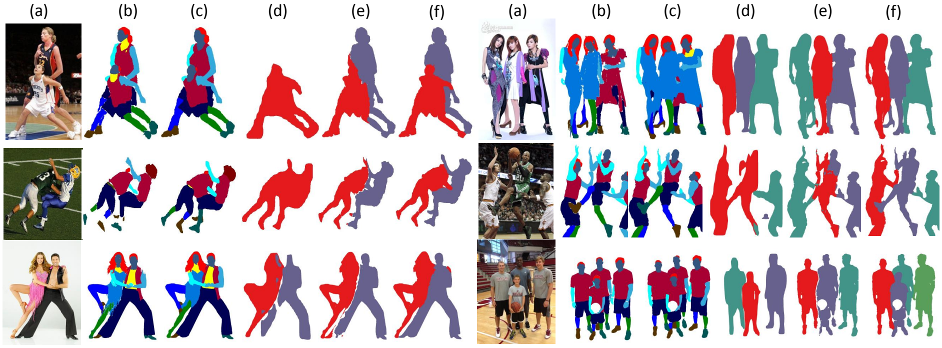

Fig. 4. Visualization of results. For each (a) input image, we show the (b) parsing ground truth, (c) global parsing prediction, person instance map predictions from (d) Mask RCNN, (e) DL and (f) MH-Parser. In (b) and (c), each color represents a semantic parsing category. In (d), (e) and (f), each color represents one person instance. We can see the proposed MH-Parser can generate satisfactory global parsing maps and person instance masks, outperforming Mask RCNN and DL when persons are closely entangled. Best viewed in the coloured pdf version with 2 times zoom.

only exploits pairwise relation between embeddings of pixels, while MH-Parser models high-order relations among different regions and shows better performance.

We note that the method Nested Adversarial Networks (NAN) in [57] and Multi-Human Parsing Machines (MHPM) [31] achieve better results on our MHP dataset. However, the superiority performance of NAN and MHPM mainly comes from using large-scale training dataset. To be specific, NAN leverages a nested deep framework containing 3 GAN models pre-trained on a large-scale dataset, which contains 25k well annotated images, to achieve high performance, and MHPM uses generated dataset (*i.e.*, 30k synthesised images) to enhance its performance. In contrast, all the results reported in this work are based on 3k training images. Based on the same training set, the comparison of NAN, MHPM and our MH-Parser is shown in Tab. 3, which can well demonstrate the superiority of our approach over the others, especially on the hard cases of close person interactions.

Table 3. Results from different methods with the same amount of training data. NAN Low-Shot refers to the NAN model trained only on the training set of the MHP dataset. MHP Solver is the model in [31], which is trained only on the training set of the MHP dataset.

	$AP^p_{0.5}$	$AP^p_{0.5}$ Top 20%	$AP^p_{0.5}$ Top 5%
NAN Low-Shot	47.82	38.53	28.99
MHP Solver	**51.07**	37.28	30.79
MH-Parser	50.10	**41.67**	**33.69**

5.3.2 Comparison with State-of-the-Arts on Separating Person Instances. We also evaluate the proposed MH-Parser on the Buffy dataset and compare it with other state-of-the-art methods. On Buffy forward score and backward score are used to evaluate the performance of person individuation [24]. We follow the same evaluation metric, and our average forward and backward scores for Episode 4, 5 and 6 on the Buffy dataset are 71.11% and 71.94%, respectively. In [24] the average forward and backward scores are 68.22% and 69.66% on the same dataset, and [52] reports an average score of 62.4%. Note that MH-Parser is not trained on Buffy, only evaluation is performed.

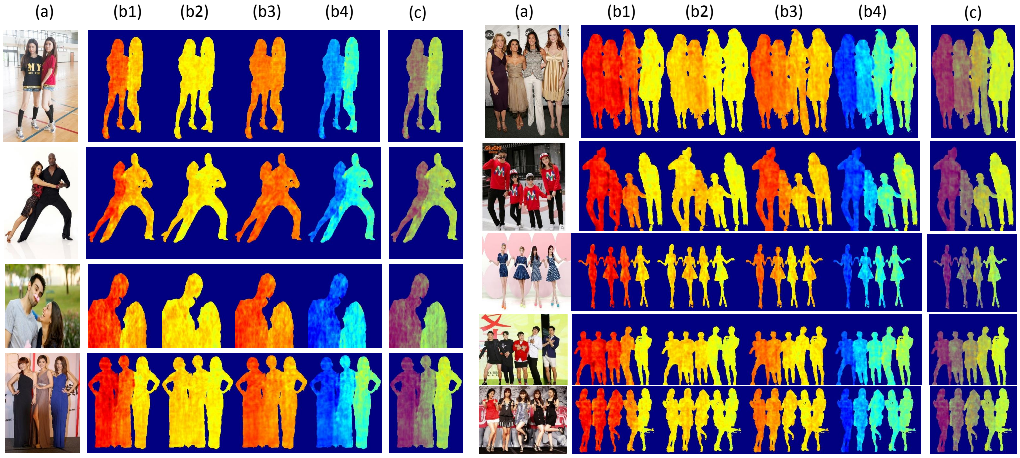

Fig. 5. Visualization of the feature maps $\hat{\mathbf{F}} \in \mathbb{R}^{H \times W \times 4}$ in the affinity prediction net (obtained by up-sampling $\mathbf{F}$ to match the spatial dimension of input images). For each (a) input image, we show (b1-4) the outputs of each channel of $\hat{\mathbf{F}}$ and (c) the superimposition of them. The feature map has 4 channels, giving a 4-dimensional accordance vector to each point of the input image. The warmer color in the feature map indicates larger accordance value in the accordance vector. When combining the information of all the 4 channels, person instances can be differentiated. The points associated with one person have similar accordance vectors (in terms of Euclidean distance) in the 4-dimension feature space and similar colors, and the points for different persons have distinct accordance vectors and different colors. Best viewed in the coloured pdf version with 2 times zoom.

We can see MH-Parser achieves the best performance compared with other state-of-the-art methods in separating closely entangled persons.

As discussed earlier, the PASCAL-Person-Part dataset is firstly used in an instance-aware setting in [32] and shares the same philosophy of separating persons and parsing person parts. This dataset contains 3,533 images, with 1,716 images for training and 1,817 images for testing. We apply our proposed method on this dataset, and report the AP^r metric over the test set for fair comparison.

The performances of our proposed MH-Parser and the methods in [32] in terms of AP^r are shown in Tab. 4. We can see that our MH-Parser outperforms the methods in [32], achieving state-of-the-art performance on this dataset. The results demonstrate the superiority of our proposed MH-Parser method. We also note that on the PASCAL-Person-Part dataset, NAN Low-Shot, which is trained on only its own training set, achieves 40.56% $AP^r_{0.5}$. For our proposed MH-Parser, it achieves 42.31% $AP^r_{0.5}$ with equal amount of training data, which shows good generalization capability.

Table 4. Results of different methods on the PASCAL-Person-Part dataset. The results of the MNC model is also taken from [32].

Method	$AP^r_{0.5}$	$AP^r_{0.6}$	$AP^r_{0.7}$	AP^r_{vol}
MNC [11]	38.80	28.10	19.30	36.70
Holistic [32]	40.60	30.40	19.10	38.40
MH-Parser	**42.31**	**34.22**	**20.14**	**40.04**

5.3.3 Components Analysis for MH-Parser. In this subsection, we test the proposed MH-Parser in various settings. All the variants of MH-Parser are trained on the MHP training set and evaluated on the validation set. The loss in Eqn. (16) is adjusted to either include or exclude the Graph-GAN term. We also demonstrate effects of the instance mask refinement. In the refinement, the pairwise term in Eqn. (21) is disabled by setting $w^{(1)}$ to 0 to investigate whether the learned pairwise term is beneficial to the refinement process. The performance of these variants in terms of AP^p, AP^p_{vol} and PCP is listed in Tab. 5.

Table 5. Results from different settings on the validation set. Refine refers to instance mask refinement, and Refine w/o PAM means in the refinement step the CRF is performed without the learned pairwise term from the pairwise affinity map.

MH-Parser	$\mathrm{AP}^p_{0.5}$	AP^p_{vol}	$\mathrm{PCP}_{0.5}$
Baseline L2	41.92	45.21	46.77
+ $\mathcal{L}_{\mathrm{GAN}}$	44.34	46.43	47.62
+ Refine, w/o PAM	49.49	48.98	50.48
+ Refine	**50.36**	**49.29**	**50.57**
w/ GT Person Number	51.39	49.77	51.32
w/ GT Affinity	55.83	51.28	55.85
w/ GT Global Seg.	91.75	77.29	82.96

From the results, we can see that compared to the $L2$ loss, the Graph-GAN can effectively improve the quality of the predicted pairwise affinity map. Better and finer affinity maps resulted from Graph-GAN help generate better grouping of the bottom level person instance information, leading to increased AP^p and PCP. The instance mask refinement, especially the learned pairwise term, plays a positive role in improving the performance of multi-human parsing.

We also use the respective ground truth annotations of the three components, *i.e.* ground truth person number, ground truth affinity graph and ground truth segmentation map, to probe the upper limits of MH-Parser in Tab. 5. We can see that the person number prediction and affinity map prediction are reasonably accurate, while the global segmentation is still the major hindrance of the problem of multi-human parsing. Improvement on global segmentation can greatly boost the performance of multi-human parsing.

5.3.4 Qualitative Comparison. Here we visually compare the results from Mask RCNN, DL and MH-Parser. The input images, global parsing ground truths, parsing predictions, predicted instance maps from Mask RCNN, DL and MH-Parser are visualized in Fig. 4. We can see that the MH-Parser captures both the fine-grained global parsing details and the information to differentiate person instances. For Mask RCNN, it has difficulties distinguishing closely entangled persons, especially when the bounding boxes of persons have large overlaps. The MH-Parser has better instance masks in such cases. MH-Parser also has better person instance masks than DL, especially at the boundary between two close instances.

We present visual results from variants of our MH-Parser in Fig. 6. From the results, we can see that the Graph-GAN in the MH-Parser helps improve the quality of instance masks over the baseline model. Also the instance mask refinement step refines the superpixel-level instance masks into pixel-level ones, giving finer grained instance parsing results.

The affinity prediction net generates the pairwise affinity map, which is constructed from its internal feature map **F**. Essentially the feature map **F** contains the information to distinguish one

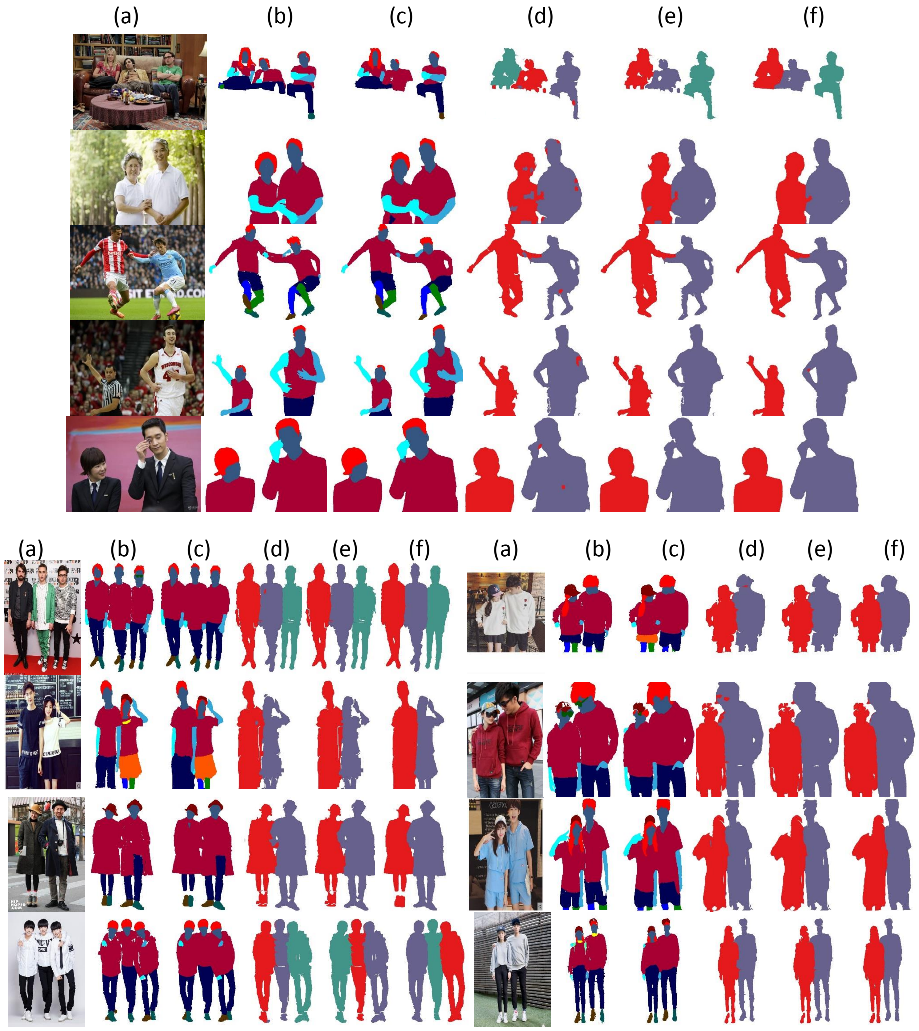

Fig. 6. Additional visualization of parsing results. For each (a) input image, we show the (b) ground truth, (c) global parsing prediction, person instance map predictions from (d) baseline $L2$ model, (e) $L2 + \mathcal{L}_{\mathrm{GAN}}$ model, and (f) $L2 + \mathcal{L}_{\mathrm{GAN}}$+Refine. In (b) and (c), each color represents a semantic parsing category. In (d), (e) and (f), each color represents one person instance. We can see the proposed MH-Parser can effectively improve the quality of person instance masks upon the baseline method. Best viewed in the coloured pdf version with 2 times zoom.

person from another in the input image. We visualize the feature map **F** in Fig. 5. Foreground masks from the corresponding parsing maps are applied to the feature maps to set the background to 0 so that foreground features can be highlighted. We can see that the feature maps assign different values to different persons in the images. Specifically, the left-most person tends to have high values in the first channel and low values in the last channel. The right most person tends to have intermediate values. The other persons are embedded with different values which roughly

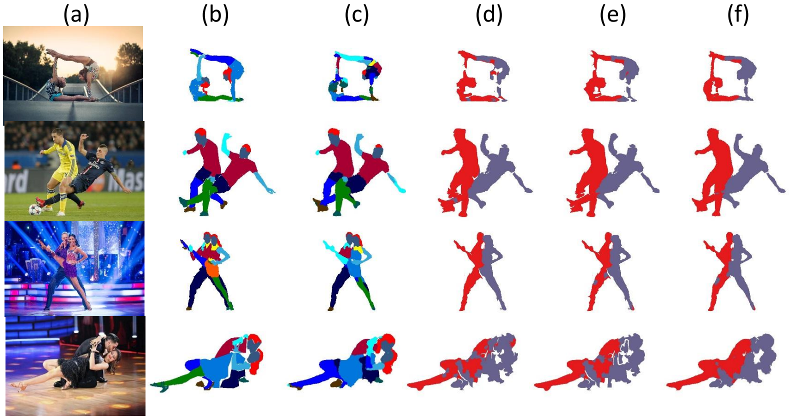

Fig. 7. Failure case analysis. Each row shows an (a) input image, (b) global parsing ground truth, (c) global parsing prediction, person instance map predictions from (d) baseline $L2$ model, (e) $L2 + \mathcal{L}_{\text{GAN}}$ model, and (f) $L2 + \mathcal{L}_{\text{GAN}}$+Refine. *Row* 1 & 2: failures due to challenging and uncommon human poses. *Row* 3 & 4: failures due to intimate interaction and occlusion between humans.

interpolate between the left and right. Thus different persons can be differentiated in the feature space.

Failure cases of some images are shown in Fig. 7, which include parsing multiple persons with occlusion, large pose variations, and interactions. The problem of multi-human parsing, reflecting realistic requirements, is very challenging and more attention is needed in the research community.

6 CONCLUSION

In this paper, we tackle the close person entanglement problem in the multi-human parsing task. With a graph-based generative adversarial model, our proposed MH-Parser is able to handle the challenging case of close person interactions and entanglement in real-world multi-human parsing scenarios. We also contributed a new multi-human parsing dataset with intensive person interactions, which aligns better with real-world scenarios. We performed detailed evaluations of the proposed method and compared with current state-of-the-art solutions to verify the effectiveness of our proposed method in addressing the challenges in multi-human parsing.

ACKNOWLEDGMENTS

The work of Jianshu Li was partially supported by National Research Foundation Singapore. The work of Jian Zhao was partially supported by **C**hina **S**cholarship **C**ouncil (CSC) grant 201503170248. The work of Jiashi Feng was partially supported by NUS startup R-263-000-C08-133, MOE Tier-I R-263-000-C21-112, NUS IDS R-263-000-C67-646 and ECRA R-263-000-C87-133.

REFERENCES

[1] Radhakrishna Achanta, Appu Shaji, Kevin Smith, Aurelien Lucchi, Pascal Fua, and Sabine Süsstrunk. 2010. *Slic superpixels.* Technical Report.

[2] Pablo Arbelaez, Michael Maire, Charless Fowlkes, and Jitendra Malik. 2011. Contour detection and hierarchical image segmentation. *The IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)* 33, 5 (2011), 898–916.

[3] Martin Arjovsky, Soumith Chintala, and Léon Bottou. 2017. Wasserstein generative adversarial networks. In *Proceedings of the International Conference on Machine Learning (ICML).* 214–223.

[4] Anurag Arnab, Sadeep Jayasumana, Shuai Zheng, and Philip HS Torr. 2016. Higher order conditional random fields in deep neural networks. In *European Conference on Computer Vision.* Springer, 524–540.

[5] Anurag Arnab and Philip HS Torr. 2017. Pixelwise instance segmentation with a dynamically instantiated network. *arXiv preprint arXiv:1704.02386* (2017).

[6] Aleksandar Bojchevski, Oleksandr Shchur, Daniel Zügner, and Stephan Günnemann. 2018. Netgan: Generating graphs via random walks. *arXiv preprint arXiv:1803.00816* (2018).

[7] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. 2017. Realtime multi-person 2d pose estimation using part affinity fields. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.* 7291–7299.

[8] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. 2016. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *arXiv preprint arXiv:1606.00915* (2016).

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

[10] Xiao Chu, Wanli Ouyang, Wei Yang, and Xiaogang Wang. 2015. Multi-task recurrent neural network for immediacy prediction. In *Proceedings of the IEEE international conference on computer vision.* 3352–3360.

[11] Jifeng Dai, Kaiming He, and Jian Sun. 2016. Instance-aware semantic segmentation via multi-task network cascades. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR).* 3150–3158.

[12] Bert De Brabandere, Davy Neven, and Luc Van Gool. 2017. Semantic instance segmentation with a discriminative loss function. *arXiv preprint arXiv:1708.02551* (2017).

[13] Michaël Defferrard, Xavier Bresson, and Pierre Vandergheynst. 2016. Convolutional neural networks on graphs with fast localized spectral filtering. In *Proceedings of the Advances in Neural Information Processing Systems (NIPS).* 3844–3852.

[14] Mark Everingham, SM Ali Eslami, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. 2015. The pascal visual object classes challenge: A retrospective. *International journal of computer vision* 111, 1 (2015), 98–136.

[15] Vittorio Ferrari, Manuel Marin-Jimenez, and Andrew Zisserman. 2008. Progressive search space reduction for human pose estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR).* 1–8.

[16] Raghudeep Gadde, Varun Jampani, Martin Kiefel, Daniel Kappler, and Peter V Gehler. 2016. Superpixel convolutional networks using bilateral inceptions. In *European Conference on Computer Vision.* Springer, 597–613.

[17] Chuang Gan, Ming Lin, Yi Yang, Gerard de Melo, and Alexander G Hauptmann. 2016. Concepts Not Alone: Exploring Pairwise Relationships for Zero-Shot Video Activity Recognition. In *Proceedings of the Association for the Advance of Artificial Intelligence Conference on Artificial Intelligence (AAAI).* 3487.

[18] Ke Gong, Xiaodan Liang, Xiaohui Shen, and Liang Lin. 2017. Look into Person: Self-supervised Structure-sensitive Learning and A New Benchmark for Human Parsing. *arXiv preprint arXiv:1703.05446* (2017).

[19] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative adversarial nets. In *Proceedings of the Advances in Neural Information Processing Systems (NIPS).* 2672–2680.

[20] Bharath Hariharan, Pablo Arbeláez, Ross Girshick, and Jitendra Malik. 2014. Simultaneous detection and segmentation. In *Proceedings of the European Conference on Computer Vision (ECCV).* 297–312.

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

[22] Rui Huang, Shu Zhang, Tianyu Li, and Ran He. 2017. Beyond Face Rotation: Global and Local Perception GAN for Photorealistic and Identity Preserving Frontal View Synthesis. *arXiv preprint arXiv:1704.04086* (2017).

[23] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. 2016. Image-to-image translation with conditional adversarial networks. *arXiv preprint arXiv:1611.07004* (2016).

[24] Hao Jiang and Kristen Grauman. 2017. Detangling People: Individuating Multiple Close People and Their Body Parts via Region Assembly. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.* 6021–6029.

[25] Diederik Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014).

[26] Thomas N Kipf and Max Welling. 2016. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907* (2016).

[27] Thomas N Kipf and Max Welling. 2016. Variational graph auto-encoders. *arXiv preprint arXiv:1611.07308* (2016).

[28] Pushmeet Kohli, Philip HS Torr, et al. 2009. Robust higher order potentials for enforcing label consistency. *International Journal of Computer Vision* 82, 3 (2009), 302–324.

[29] Philipp Krähenbühl and Vladlen Koltun. 2011. Efficient inference in fully connected crfs with gaussian edge potentials. In *Proceedings of the Advances in Neural Information Processing Systems (NIPS).* 109–117.

[30] John Lafferty, Andrew McCallum, and Fernando CN Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the International Conference on Machine Learning (ICML).*

[31] Jianshu Li, Jian Zhao, Yunpeng Chen, Sujoy Roy, Shuicheng Yan, Jiashi Feng, and Terence Sim. 2018. Multi-human parsing machines. In *Proceedings of the 26th ACM international conference on Multimedia.* 45–53.

[32] Qizhu Li, Anurag Arnab, and Philip HS Torr. 2017. Holistic, Instance-Level Human Parsing. *arXiv preprint arXiv:1709.03612* (2017).

[33] Yi Li, Haozhi Qi, Jifeng Dai, Xiangyang Ji, and Yichen Wei. 2016. Fully Convolutional Instance-aware Semantic Segmentation. *arXiv preprint arXiv:1611.07709* (2016).

[34] Yujia Li, Daniel Tarlow, Marc Brockschmidt, and Richard Zemel. 2015. Gated graph sequence neural networks. *arXiv preprint arXiv:1511.05493* (2015).

[35] Xiaodan Liang, Si Liu, Xiaohui Shen, Jianchao Yang, Luoqi Liu, Jian Dong, Liang Lin, and Shuicheng Yan. 2015. Deep human parsing with active template regression. *IEEE transactions on pattern analysis and machine intelligence* 37, 12 (2015), 2402–2414.

[36] Xiaodan Liang, Xiaohui Shen, Donglai Xiang, Jiashi Feng, Liang Lin, and Shuicheng Yan. 2016. Semantic object parsing with local-global long short-term memory. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR).* 3185–3193.

[37] Xiaodan Liang, Yunchao Wei, Xiaohui Shen, Jianchao Yang, Liang Lin, and Shuicheng Yan. 2015. Proposal-free network for instance-level object segmentation. *arXiv preprint arXiv:1509.02636* (2015).

[38] Xiaodan Liang, Chunyan Xu, Xiaohui Shen, Jianchao Yang, Si Liu, Jinhui Tang, Liang Lin, and Shuicheng Yan. 2015. Human parsing with contextualized convolutional neural network. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV).* 1386–1394.

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

[40] Zhouhan Lin, Minwei Feng, Cicero Nogueira dos Santos, Mo Yu, Bing Xiang, Bowen Zhou, and Yoshua Bengio. 2017. A structured self-attentive sentence embedding. *arXiv preprint arXiv:1703.03130* (2017).

[41] Si Liu, Xiaodan Liang, Luoqi Liu, Xiaohui Shen, Jianchao Yang, Changsheng Xu, Liang Lin, Xiaochun Cao, and Shuicheng Yan. 2015. Matching-cnn meets knn: Quasi-parametric human parsing. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR).* 1419–1427.

[42] Pauline Luc, Camille Couprie, Soumith Chintala, and Jakob Verbeek. 2016. Semantic segmentation using adversarial networks. *arXiv preprint arXiv:1611.08408* (2016).

[43] Franco Manessi, Alessandro Rozza, and Mario Manzo. 2017. Dynamic Graph Convolutional Networks. *arXiv preprint arXiv:1704.06199* (2017).

[44] Davy Neven, Bert De Brabandere, Stamatios Georgoulis, Marc Proesmans, and Luc Van Gool. 2017. Fast Scene Understanding for Autonomous Driving. *arXiv preprint arXiv:1708.02550* (2017).

[45] Alejandro Newell, Zhiao Huang, and Jia Deng. 2016. Associative Embedding: End-to-End Learning for Joint Detection and Grouping. *arXiv preprint arXiv:1611.05424* (2016).

[46] Zhang Ning, Paluri Manohar, Taigman Yaniv, Fergus Rob, and Bourdev Lubomir. 2015. Beyond Frontal Faces: Improving Person Recognition Using Multiple Cues. arXiv:arXiv:1501.05703

[47] Alec Radford, Luke Metz, and Soumith Chintala. 2015. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434* (2015).

[48] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Proceedings of the Advances in Neural Information Processing Systems (NIPS).* 91–99.

[49] Girshick Ross, Radosavovic Ilija, Gkioxari Georgia, Dollár Piotr, and He Kaiming. 2018. Detectron. https://github.com/facebookresearch/detectron.

[50] Chris Russell, Pushmeet Kohli, Philip HS Torr, et al. 2009. Associative hierarchical crfs for object class image segmentation. In *2009 IEEE 12th International Conference on Computer Vision (ICCV).* IEEE, 739–746.

[51] Benjamin Sapp and Ben Taskar. 2013. MODEC: Multimodal Decomposable Models for Human Pose Estimation. In *In Proc. CVPR.*

[52] Vibhav Vineet, Jonathan Warrell, Lubor Ladicky, and Philip HS Torr. 2011. Human Instance Segmentation from Video using Detector-based Conditional Random Fields.. In *BMVC*, Vol. 2. 12–15.

[53] S Vichy N Vishwanathan, Nicol N Schraudolph, Risi Kondor, and Karsten M Borgwardt. 2010. Graph kernels. *Journal of Machine Learning Research* 11, Apr (2010), 1201–1242.

[54] Hongwei Wang, Jia Wang, Jialin Wang, Miao Zhao, Weinan Zhang, Fuzheng Zhang, Xing Xie, and Minyi Guo. 2018. GraphGAN: graph representation learning with generative adversarial nets. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

[55] Kota Yamaguchi, M Hadi Kiapour, Luis E Ortiz, and Tamara L Berg. 2012. Parsing clothing in fashion photographs. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 3570–3577.

[56] Zhang Zhanpeng, Luo Ping, Chen Change Loy, and Tang Xiaoou. 2016. From Facial Expression Recognition to Interpersonal Relation Prediction. In *arXiv:1609.06426v2*.

[57] Jian Zhao, Jianshu Li, Yu Cheng, Terence Sim, Shuicheng Yan, and Jiashi Feng. 2018. Understanding humans in crowded scenes: Deep nested adversarial learning and a new benchmark for multi-human parsing. In *Proceedings of the 26th ACM international conference on Multimedia*. 792–800.