Multi-Human Parsing Machines

Jianshu Li [1,3] Jian Zhao [2] Yunpeng Chen [2]
Sujoy Roy [3] Shuicheng Yan [2] Jiashi Feng [2] Terence Sim [1]
[1] School of Computing, National University of Singapore
[2] Electrical & Computer Engineering, National University of Singapore
[3] SAP Machine Learning Singapore
{jianshu,zhaojian90,chenyunpeng}@u.nus.edu,sujoy.roy@sap.com
{elefjia,eleyans}@nus.edu.sg,tsim@comp.nus.edu.sg

ABSTRACT

Human parsing is an important task in human-centric analysis. Despite the remarkable progress in single-human parsing, the more realistic case of multi-human parsing remains challenging in terms of the data and the model. Compared with the considerable number of available single-human parsing datasets, the datasets for multi-human parsing are very limited in number mainly due to the huge annotation effort required. Besides the data challenge to multi-human parsing, the persons in real-world scenarios are often entangled with each other due to close interaction and body occlusion, making it difficult to distinguish body parts from different person instances. In this paper we propose the Multi-Human Parsing Machines (MHPM) system, which contains an MHP Montage model and an MHP Solver, to address both challenges in multi-human parsing. Specifically, the MHP Montage model in MHPM generates realistic images with multiple persons together with the parsing labels. It intelligently composes single persons onto background scene images while maintaining the structural information between persons and the scene. The generated images can be used to train better multi-human parsing algorithms. On the other hand, the MHP Solver in MHPM solves the bottleneck of distinguishing multiple entangled persons with close interaction. It employs a Group-Individual Push and Pull (GIPP) loss function, which can effectively separate persons with close interaction. We experimentally show that the proposed MHPM can achieve state-of-the-art performance on the multi-human parsing benchmark and the person individualization benchmark, which distinguishes closely entangled person instances.

CCS CONCEPTS

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

MM '18, October 22–26, Seoul, Republic of Korea, 2018

ACM ISBN 978-1-4503-5665-7/18/10…$15.00
https://doi.org/10.1145/3240508.3240515

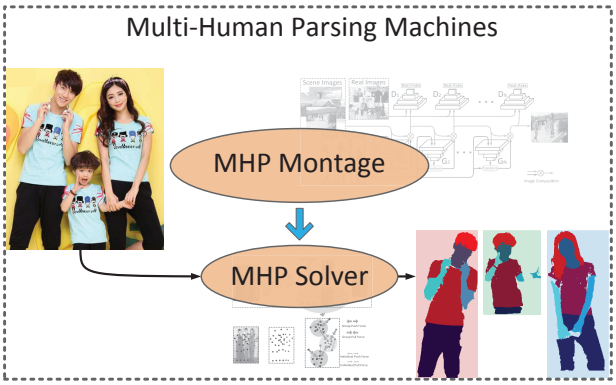

Figure 1: Overview of the proposed Multi-Human Parsing Machines (MHPM) system. The MHPM consists of an MHP Montage model, which automatically generates training data for multi-human parsing without costing any laborious human annotation effort, and also an MHP Solver, which uses a simple yet effective loss to distinguish person instances with close interaction. The proposed MHPM provides a one-stop solution to address jointly data challenge and model challenge in multi-human parsing.

KEYWORDS

Human Parsing; Multi-Human Parsing; Human-Centric Image Analysis; Generative Adversarial Networks; Image Composition; Instance Segmentation

ACM Reference Format:
Jianshu Li, Jian Zhao, Yunpeng Chen, Sujoy Roy, Shuicheng Yan, Jiashi Feng, Terence Sim. 2018. Multi-Human Parsing Machines. In 2018 ACM Multimedia Conference (MM'18), October 22–26, 2018, Seoul, Republic of Korea. ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3240508.3240515

1 INTRODUCTION

Multi-human parsing refers to simultaneously partitioning the persons in an image into multiple semantically consistent regions and individualizing different person instances. This task is much closer to real world scenarios compared to the widely researched human parsing task, which only considers one person at a time, thus it is attracting growing research attention recently.

Towards solving this practical multi-human parsing problem, there are mainly two challenges. The first is the lack of available data resources, which we term as the *data challenge*. To date, the number and size of multi-human parsing datasets are rather limited, due to the huge annotation effort required. In [13], a multi-human parsing dataset was proposed but only contains 5,000 images. In comparison, for single-human parsing, there are a number of large-scale datasets containing tens of thousands images. For example, the ATR [17] dataset contains 17,700 images and the Look Into Person [6] dataset contains 50,462 image. Rich data resources for single-human parsing have greatly boosted the performance of single-human parsing models, while multi-human parsing is severely challenged by scarcity of carefully annotated data. The second challenge is it is difficult to separate closely entangled persons in multi-human parsing, which is called the *model challenge* in this work. Although multi-human parsing can be straightforwardly solved by applying a person detector and a single-human parser sequentially, standard person detectors only work well for upright persons, *e.g.* pedestrians, with simple poses. However, in real-world scenarios, the persons do not always position themselves upright. Close interaction, body occlusion and entanglement are very common in human-centric images. How to effectively separate these person instances remains a big difficulty for multi-human parsing models and also for other human-centric images analysis applications.

In this paper, targeting at the data challenge and the model challenge, we propose a unified multi-human parsing system, named Multi-Human Paring Machines (MHPM) as illustrated in Fig. 1. In particular, to tackle the data challenge, we design an MHP Montage model as the data generator in MHPM. This model can intelligently compose images from existing datasets to generate realistic images containing multiple persons together with their annotations. It uses a novel image composition network which employs Generative Adversarial Networks (GAN) [7] and Spatial Transformer Networks (STN) [11] to learn to automatically compose images. In this way, the MHP Montage model can effortlessly generate new training data to train multi-human parsing models, and therefore effectively solves the data challenge in the multi-human parsing task.

To address the model challenge in multi-human parsing, the proposed MHPM also contains a multi-human parser, called the MHP Solver. The MHP Solver transforms input images from the raw pixel space into an embedding space, where the embeddings of the same person are close to each other while those of different persons are far away. To effectively learn the mapping from the raw pixel space to the desired embedding space, the MHP Solver employs a novel Global-Individual Push and Pull (GIPP) loss, which operates on the embedding space. It contains two levels of embedding learning losses, *i.e.* a group level loss operating on the centers of embedding clusters and an individual level loss operating on embeddings of each pixel. Thus with the GIPP loss, the MHP Solver can effectively distinguish the closely entangled person instances.

Our contributions are mainly three-fold. 1) We propose an MHP Montage model to automatically synthesize realistic data with annotations for facilitating multi-human parsing model training. 2) We propose an MHP Solver with a novel loss to distinguish closely entangled persons. 3) The proposed Multi-Human Parsing Machines system provides a one-stop solution to the challenging multi-human parsing task.

2 RELATED WORK

In this section, we briefly review previous methods related to our work, including human parsing, instance segmentation and image composition with Generative Adversarial Networks (GAN).

Human Parsing. Most of the previous human parsing methods [6, 16, 17, 19, 27] and datasets [3, 15, 17, 25] focus on single-human parsing. Only recently, [13] first proposed the more realistic multi-human parsing problem, taking into account the simultaneous presence of multiple persons in an instance-aware setting with occlusion and interaction between persons. Although multi-human parsing aligns much better with reality, there are much fewer available datasets compared to single-human parsing, and multi-human parsing is harder than the single-human setting. Our proposed MHPM tackles multi-human parsing based on the two aspects of challenges, making a step towards solving this problem.

Instance Segmentation. The multi-human parsing problem is related to the instance-aware semantic segmentation problem. Instance-aware human segmentation only provides the silhouette of each person instance, while multi-human parsing gives more details within each person instance. For instance segmentation, there are two main streams of methods, top-down methods and bottom-up methods. The top-down methods [4, 8, 14] usually use a detection component to localize each instance, which is further processed to generate pixel segmentation. They heavily depend on the performance of the detection component, and suffer performance drop when the detection component does not work properly *e.g.* in real-world human parsing scenarios where different person instances are close to each other. The bottom-up methods [5, 13, 20] do not rely on detection components. They convert the input images from pixel space to embedding space, where different instances can be identified. Different from the existing bottom-up methods, the proposed MHPM uses a novel grouping loss to guide the learning of the embedding space.

Image Blending and Composition with Generative Adversarial Networks. GAN-based methods [1, 7] focus on generating photo-realistic images [10] from scratch. Using GANs to guide the image blending and composition process is also explored recently. In [24], high-resolution well-blended images are generated given copy-and-paste ones by using GANs to adjust the color tones of the foreground images so that they can better blend into the background. Similarly, image blending is performed with end-to-end supervised learning in [22]. These image blending methods only learn to adjust the color of foreground images, and do not consider where to put the foreground images. In [18], spatial transformer networks are used to find geometric corrections to foreground images such that the composed images are natural and realistic. Different from these existing works, our MHP Montage model uses a sequence of networks which compose foreground images onto background in a sequential fashion with conditional dependencies to compose multiple foreground images, with an ultimate goal of benefiting learning of multi-human parsing algorithms.

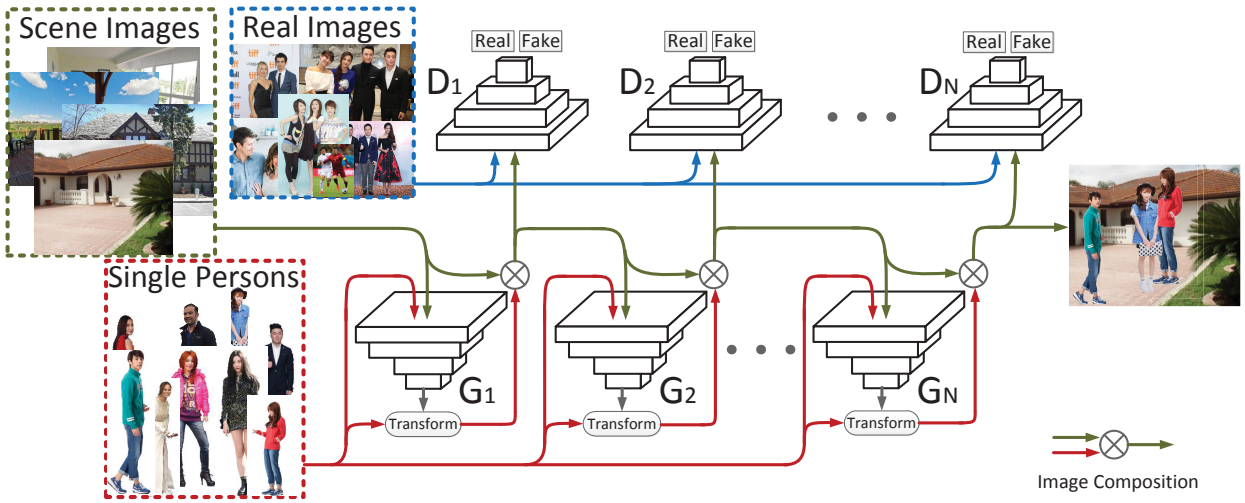

Figure 2: Overall structure of the proposed MHP Montage model in MHPM. The neural networks in the top row are discriminators and those in the bottom row are generators. The arrows in different colors denote the information flow of different groups of inputs: blue arrows represent real-world images containing multiple persons, dark green ones denote real-world scene images, and red arrows denote single-human images. The circled cross symbol denotes image composition operation defined in Eqn. (5)**.**

3 MULTI-HUMAN PARSING MACHINES

In this section, we elaborate on the proposed Multi-Human Parsing Machines (MHPM) system. We first introduce the MHP Montage model and then the MHP Solver.

3.1 MHP Montage Model

The overall structure of the MHP Montage model is shown in Fig. 2. The MHP Montage model takes as input triplets of images, including 1) real-world images with multiple persons (without any annotations), 2) real-world scene images which do not contain any persons, and 3) images that only contain single person instances without any background. It learns how to compose the single person images, onto the scene images, such that the resultant composed images are as realistic as natural ones with multiple person instances. The real-world images with multiple persons can be easily obtained from datasets for human-centric tasks such as people recognition in photo album [21], interpersonal relation prediction [26], and many other on-line images obtained by querying web search engines. The real-world scene images are also widely available in datasets for scene understanding, such as Places [28], ADE20K [29, 30], *etc.* For the last group, the images with only person instances can also be easily obtained from human parsing datasets, where parsing masks are available to remove the existing background pixels and only pixels belonging to the foreground persons remain.

The MHP Montage model consists of a cascade of generators $G_1, G_2, \cdots, G_N$, and a cascade of corresponding discriminators $D_1, D_2, \cdots, D_N$. Different generators form a cascaded structure, *i.e.* the output from one generator is the input of the following generator. Each generator G_i takes as input one background image $\mathbf{I}_{\text{bg}}^{i-1}$ (dark green arrows in Fig. 2) and one foreground image $\mathbf{I}_{\text{fg}}^{i}$ together with its corresponding binary foreground mask $\mathbf{M}_{\text{fg}}^{i}$ (red arrows in Fig. 2), and produces a set of transformation parameters θ_i:

$$\theta^i = G_i(\mathbf{I}_{\text{bg}}^{i-1}, \mathbf{I}_{\text{fg}}^{i}, \mathbf{M}_{\text{fg}}^{i}). \tag{1}$$

The parameters θ_i are used to transform the foreground image $\mathbf{I}_{\text{fg}}^{i}$ and its mask $\mathbf{M}_{\text{fg}}^{i}$ into a new image $\hat{\mathbf{I}}_{\text{fg}}^{i}$ and a new mask $\hat{\mathbf{M}}_{\text{fg}}^{i}$:

$$\begin{aligned} \hat{\mathbf{I}}_{\text{fg}}^{i} &= T_{\theta^i}(\mathbf{I}_{\text{fg}}^{i}), \\ \hat{\mathbf{M}}_{\text{fg}}^{i} &= T_{\theta^i}(\mathbf{M}_{\text{fg}}^{i}). \end{aligned} \tag{2}$$

The transformations are modeled by Spatial Transformer Networks (STN) [11], and we parametrize the transform parameters as

$$\theta^i = \begin{bmatrix} \theta_{11}^i & \theta_{12}^i & d_x \\ \theta_{21}^i & \theta_{22}^i & d_y \end{bmatrix}, \tag{3}$$

which models affine transformations. We find this kind of transformation is enough for generating useful training data for multi-human parsing. The transformation T_{θ^i} is realized by

$$\begin{bmatrix} x \\ y \end{bmatrix} = \theta^i \begin{bmatrix} \hat{x} \\ \hat{y} \\ 1 \end{bmatrix}, \tag{4}$$

where x and y are the row and column indices on the input to the transformation operation T_{θ_i}, and $\hat{x}$ and $\hat{y}$ are the row and column indices on the output from the transformation. Specifically, the pixel values on the output $\hat{\mathbf{I}}_{\text{fg}}^{i}$ are obtained pixel by pixel from the input $\mathbf{I}_{\text{fg}}^{i}$, *i.e.* the pixel value at location $(\hat{x}, \hat{y})$ on $\hat{\mathbf{I}}_{\text{fg}}^{i}$ is obtained by interpolating the pixel values on $\mathbf{I}_{\text{fg}}^{i}$ centered at (x, y) in a bilinear fashion. Similarly, the mask $\mathbf{M}_{\text{fg}}^{i}$ is also transformed using the same parameter θ^i into $\hat{\mathbf{M}}_{\text{fg}}^{i}$.

With the transformed foreground image $\hat{\mathbf{I}}^i_{\text{fg}}$ and mask $\hat{\mathbf{M}}^i_{\text{fg}}$, we can obtain the composed image pixel-wisely:

$$\begin{aligned}\mathbf{I}^i_{\text{bg}} =& \mathbf{I}^{i-1}_{\text{bg}} \otimes \hat{\mathbf{I}}^i_{\text{fg}} \\ =& \mathbf{I}^{i-1}_{\text{bg}}(1 - \hat{\mathbf{M}}^i_{\text{fg}}) + \hat{\mathbf{I}}^i_{\text{fg}}\hat{\mathbf{M}}^i_{\text{fg}}.\end{aligned} \tag{5}$$

Here $\otimes$ denotes the composition operation, which superimposes the foreground image onto the background images, *i.e.* the pixels in the background images are replaced with the corresponding pixels within the foreground persons. The composition operation is equivalent to the summation of the pixel-wise multiplication between $\hat{\mathbf{I}}^i_{\text{fg}}$ and $\hat{\mathbf{M}}^i_{\text{fg}}$ and the pixel-wise multiplication between $\mathbf{I}^{i-1}_{\text{bg}}$ and $(1 - \hat{\mathbf{M}}^i_{\text{fg}})$.

Putting them together, we can write the generated image $\mathbf{I}^i_{\text{bg}}$ as a function of the generator G_i and the inputs as

$$\mathbf{I}^i_{\text{bg}} = f_{G_i}(\mathbf{I}^{i-1}_{\text{bg}}, \mathbf{I}^i_{\text{fg}}, \mathbf{M}^i_{\text{fg}}). \tag{6}$$

Each discriminator D_i takes as input an image, either $\mathbf{I}^i_{\text{bg}}$ or a real world multi-human image $\mathbf{I}^i_{\text{mh}}$ (blue arrows in Fig. 2). D_i outputs a real value $D_i(\mathbf{I})$ between 0 and 1, which indicates the probability of the input image being a real-world multi-human image. The discriminator D_i is responsible for distinguishing the composed image I^i_{bg} from the generator G_i, thus there are equal numbers of generators and discriminators in the MHP Montage model.

To train the MHP Montage model, the generators and discriminators play a minmax game with the following objective:

$$\mathcal{L}_{\text{GAN}} = \sum_i^N \log\left(D_i(\mathbf{I}^i_{\text{mh}})\right) + \log\left(1 - D_i(f_{G_i}(\mathbf{I}^{i-1}_{\text{bg}}, \mathbf{I}^i_{\text{fg}}, \mathbf{M}^i_{\text{fg}}))\right), \tag{7}$$

and we aim to find the optimal $G^*_i, \forall i = 1, 2, \cdots, N$ such that

$$G^* = \arg \min_{\{\text{G}_1, \cdots, \text{G}_\text{N}\}} \max_{\{\text{D}_1, \cdots, \text{D}_\text{N}\}} \mathcal{L}_{\text{GAN}}(\text{G}_1, \cdots, \text{G}_\text{N}, \text{D}_1, \cdots, \text{D}_\text{N}). \tag{8}$$

Specifically, $\mathbf{I}^0_{\text{bg}}$ is an image drawn from the scene image pool, $\mathbf{I}^1_{\text{bg}}$ is the composed image which has one synthetic person in the scene image, $\mathbf{I}^2_{\text{bg}}$ is the composed image with two synthetic persons, and $\mathbf{I}^N_{\text{bg}}$ is the final output from the MHP Montage model with N synthetic persons. The generation of each additional foreground person is conditioned on the original scene image as well as the already added (composed) foreground images. Each generator aims at finding the best transformation of the foreground image, such that the composed images look natural and realistic. The discriminators aim to distinguish whether the input image is a real image or a composed (fake) one. During the adversarial training, the discriminators guide the generation process of the generator, such that the generated images look like real ones as much as possible.

After training the MHP Montage model, it can be used to generate multi-human images with up to N persons. We choose the best N^* such that

$$N^* = \arg \max_{\text{n}=1, \cdots, \text{N}} \text{D}_\text{n}(\mathbf{I}^\text{n}_{\text{bg}}). \tag{9}$$

The generated image $\mathbf{I}^{N^*}_{\text{bg}}$ contains N^* persons. When generating $\mathbf{I}^{N^*}_{\text{bg}}$, the transformations from each generator can be used to simultaneously transform the corresponding semantic segmentation annotations of the single-human images. Therefore when a new multi-human image is synthesized, the corresponding multi-human annotation is also synthesized from the single-human annotations and the transformation parameters $\theta_i, \forall i = 1, \cdots, N^*$. Thus the obtained multi-human image and annotation pair can be used to train any multi-human parsing algorithm in a supervised manner.

3.2 MHP Solver

We then introduce the proposed MHP Solver, which aims to address the model challenge in multi-human parsing. The overall structure of the MHP Solver is shown in the upper panel of Fig. 3.

The proposed MHP Solver uses a multi-task learning strategy to tackle the multi-human parsing problem. Particularly, it uses a neural network model with one trunk network and two branches of subnetworks. The trunk network is for learning common features for the following two subnetworks, and the two subnetworks are for instance-agnostic global human parsing and instance segmentation of persons, respectively. Specifically, the global human parsing subnetwork takes as input the common feature from the trunk network, and produces the pixel-wise global parsing map of the input image. The function of this subnetwork is similar to the traditional human parsing models and other semantic segmentation models. Different from traditional human parsing models, the MHP Solver has an instance segmentation subnetwork to distinguish different person instances. The instance segmentation subnetwork also uses the common feature from the trunk network as input, and generates a set of C-dimensional embeddings for each pixel in the input image. The pixel-wise embedding vectors serve as tags for different persons, such that the embedding vectors for the same person instance are close to each other, and those for different person instances are far away from each other.

To train the MHP Solver, we use the following loss function

$$\mathcal{L} = \mathcal{L}_{\text{parsing}} + \mathcal{L}_{\text{tag}}, \tag{10}$$

which contains $\mathcal{L}_{\text{parsing}}$ for the global human parsing subnetwork and $\mathcal{L}_{\text{tag}}$ for the instance segmentation subnetwork. Specifically, the global human parsing loss is a traditional pixel-wise cross-entropy loss. Given an input colored image $\mathbf{I} \in \mathbb{R}^{H \times W \times 3}$, the output of the parsing subnetwork $\mathbf{S} \in \mathbb{R}^{H \times W \times C_0}$ with C_0 semantic parsing categories is directly supervised by the corresponding parsing label.

To effectively learn the desired pixel-wise embedding, we design a new loss function called Group-Individual Push and Pull (GIPP) loss, as illustrated in the lower panel of Fig. 3. Specifically, the GIPP loss consists of four components:

$$\mathcal{L}_{\text{tag}} = \mathcal{L}^G_{\text{push}} + \mathcal{L}^G_{\text{pull}} + \mathcal{L}^I_{\text{push}} + \mathcal{L}^I_{\text{pull}}. \tag{11}$$

According to the functionality, the four components in GIPP can be categorized into two types, the push loss and the pull loss. The push loss is incurred when the embeddings of different person instances are too close to each other, and the pull loss is incurred when the embeddings of the same person instance are too far away from each other. The push loss encourages embeddings of pixels from different person instances to be far away from each other, and the pull loss, on the other hand, encourages embeddings of pixels from the same person to be close to each other.

In GIPP, there are two types of push and pull losses, *i.e.* group level ones and individual level ones. All the four components in $\mathcal{L}_{\text{tag}}$

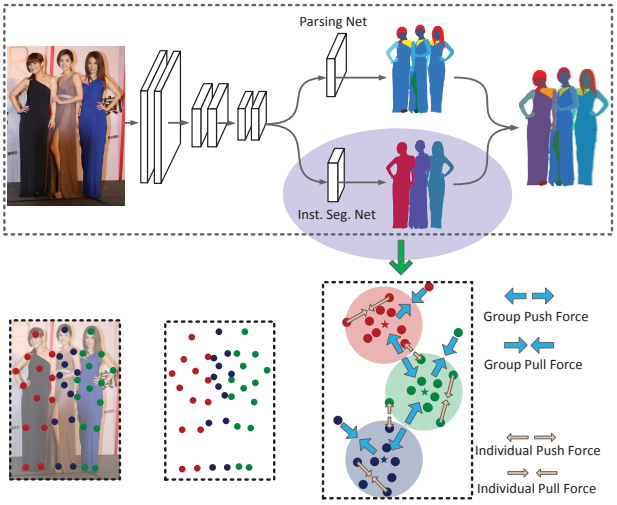

Figure 3: Illustration of the proposed MHP Solver (upper panel) with GIPP loss (lower panel) for multi-human parsing. In the lower panel, the dots of the same color represent all pixels belonging to the same person instance. The GIPP loss guides the learning of embeddings of different person instances to form different clusters in the embedding space.

operate on the embedding vectors from the output of the instance segmentation net. Given an input colored image $\mathbf{I} \in \mathbb{R}^{H\times W\times 3}$, the embedding from the instance segmentation subnetwork is $\mathbf{E} \in \mathbb{R}^{H\times W\times C}$. Each pixel $\mathbf{I}_i$ in the input image space corresponds to an embedding vector $\mathbf{E}_i \in \mathbb{R}^C$ in the embedding space. During training, the labels provide the information on the distance between the embeddings of all the pixels. Suppose there are P persons in the image, and the p-th person takes N_p pixels in the input image. Let $\mathbf{E}^p$ denote all the embedding vectors belonging to the p-th person, *i.e.* $\mathbf{E}^p \triangleq \{\mathbf{E}_i\}\forall i \in 1, 2, \cdots, N_p$. The center of $\mathbf{E}^p$ is denoted by $\mu_p = \frac{1}{N_p}\sum_{i=1}^{N_p}\mathbf{E}_i$. The group level push loss operates on centers of embeddings $\mu_p, \forall p = 1, 2, \cdots, P$:

$$\mathcal{L}^G_{\text{push}} = \frac{2}{P(P-1)}\sum_{p=1}^{P}\sum_{p'=p+1}^{P} f_{\text{push}}(\mu_p, \mu'_p). \tag{12}$$

Here the push force $f_{\text{push}}(\cdot,\cdot)$ ensures that the inputs are far away from each other by penalizing a small distance between the inputs with a hinge loss:

$$f_{\text{push}}(\mathbf{e}_1, \mathbf{e}_2) = \min(0, \delta_{\text{push}} - \|\mathbf{e}_1 - \mathbf{e}_2\|)^2, \tag{13}$$

where $\mathbf{e}_1$ and $\mathbf{e}_2$ are two input embedding vectors, δ_{push} is the minimal distance within which the push force takes effects, and $\|\cdot\|$ denotes the Euclidean distance. In other words, f_{push} aims to push the input embedding vectors to be at least as far as δ_{push}. The group level push loss ensures that the distances between any two centers of the embeddings of different persons are far away.

On the other hand, the group level pull force operates on the center embedding of one person instance and all the embeddings of this person instance:

$$\mathcal{L}^G_{\text{pull}} = \frac{1}{P}\sum_{p=1}^{P}\frac{1}{N_p}\sum_{i=1}^{N_p} f_{\text{pull}}(\mu_p, \mathbf{E}^p_i). \tag{14}$$

Here the pull force $f_{\text{pull}}(\cdot,\cdot)$ ensures that the inputs are near to each other by penalizing a large distance between the inputs with a hinge loss:

$$f_{\text{pull}}(\mathbf{e}_1, \mathbf{e}_2) = \min(0, \|\mathbf{e}_1 - \mathbf{e}_2\| - \delta_{\text{pull}})^2, \tag{15}$$

where δ_{pull} is the maximal distance beyond which the pull force takes effects. In contrast to the push force, f_{pull} aims to pull the input embedding vectors to be at most as far as δ_{pull}. The group level pull loss ensures that the distance between any embedding vector and its respective center embedding is close.

The individual losses operate on each of the individual embeddings rather than the center embedding. The individual level push loss is defined as

$$\mathcal{L}^I_{\text{push}} = \frac{2}{P(P-1)}\sum_{p=1}^{P}\sum_{p'=p+1}^{P}\sum_{i=1}^{N_p}\sum_{j=1}^{N_{p'}} f_{\text{push}}(\mathbf{E}^p_i, \mathbf{E}^{p'}_j). \tag{16}$$

The $\mathcal{L}^I_{\text{push}}$ is used to refine the individual embedding near the boundary between two adjacent persons in the embedding space such that embeddings belonging to different person instances stay far away.

For the individual level pull loss, it is defined as

$$\mathcal{L}^I_{\text{pull}} = \frac{1}{P}\sum_{p=1}^{P}\frac{1}{N_p^2}\sum_{i=1}^{N_p}\sum_{j=1}^{N_p} f_{\text{pull}}(\mathbf{E}^p_i, \mathbf{E}^p_j). \tag{17}$$

It is used to shrink the internal space of the embedding vectors for each person instance in a pairwise fashion, such that the embedding vectors of each person are close to each other, resulting in more compact embeddings in the embedding space. During implementation, we find that randomly sampling K pixels in each instance can reduce the complexity of the individual loss, and the performance is similar to the case of no sampling. We use $K = 100$ in all cases.

With all the losses defined above, the MHP Solver can be trained end-to-end with labeled images, either from a multi-human parsing dataset or from the synthetic data generated by the MHP Montage model. By controlling the margins by setting $\delta_{\text{push}} > \delta_{\text{pull}}$, we can enforce the embedding vectors of one person instance to be closer to the embeddings of the same person than to the embeddings of other persons. During training, the margins can be set to be more stringent, *i.e.* to push even further and to pull even closer, to learn more compact embeddings. During inference, we cluster the embedding vectors with the mean-shift algorithm in [5] to obtain person instances.

4 EXPERIMENTS

We conduct experiments to validate the efficacy of MHPM on the multi-human parsing task and the person individualization task (*i.e.* instance segmentation of closely entangled persons). Some ablation study of MHPM is also performed in this section.

4.1 Experimental Setting

4.1.1 Datasets. We mainly conduct experiments on the multi-human parsing benchmark MHP from [13] and the person individualization dataset Buffy from [12, 23].

In multi-human parsing, pixel level predictions of all pixels in the image into different semantic categories as well as different person instances are required. The MHP dataset [13] contains 4,980 images with such kinds of annotations. Specifically, the MHP dataset contains a training set of 3,000 images, a validation set of 1,000 images and a test set of 980 images. The other benchmark, the Buffy, is a small one with 748 images in total. Thus we only perform testing on it without training or finetuning.

For the proposed MHP Montage model, we use the Scene Parsing dataset [30] derived from ADE20K Dataset [29], which contains more than 20K scene-centric images exhaustively annotated with objects and object parts. Specifically, we use the images in the training and validation sets as the initial background pool in the MHP Montage model. The background pool is filtered to remove the images with larger side smaller than 400 pixels, or those already containing one or more person instances. The filtered background pool contains about 10K images. For the real-word multi-human images, we directly use the training images (without annotation information) in the MHP dataset from [13]. For the single-human images, we either use the ATR dataset [17], or the single person instances derived from the MHP dataset.

4.1.2 Implementation Details. In the proposed MHP Montage model, each generator is a small network with 5 convolution layers with RELU activation after each layer. The input images are resized to 144×144. An average pooling is used after the fifth convolution, followed by two fully connected layers to output the transformation parameters. The transformations are modeled by affine transformations. We use 8 generators, such that at most 8 person can be synthesized. Each generator has its own parameters. There are also 8 discriminators, which have the same depth as the generators but use leaky RELU with slope of 0.2 as the activation function. All the 8 discriminators share the same set of parameters, as it is relatively easy to distinguish synthetic images and real images.

In the proposed MHP Solver, we adopt the model from Deeplab-v2 [2] as the backbone network, which is a residual network [9] with 101 layers (ResNet-101). The first four stages in ResNet-101 are used in the trunk network, and the fifth stage and the final classifier are used in the global parsing subnetwork and the instance segmentation subnetwork. The MHP Solver is pretrained on the global human parsing task, and then it is trained on both the global human parsing and human instance segmentation tasks jointly. As input to the network, the shorter side of each input image is resized to 600 pixels, while keeping the longer side no larger than 1,000 pixels. We use $\delta_{\text{push}} = 3.0, \delta_{\text{pull}} = 0.5$ and $C = 8$ in the GIPP loss. Both the pretraining and the training are performed for 20 epoches with an exponentially decayed learning rate with a power of 0.9. The optimizer is stochastic gradient decent with momentum of 0.9, and the initial learning rates for pretraining and training are 2.5×10^{-4} and 1×10^{-5}, respectively. Standard CRF post-processing is applied to the global parsing maps with parameters from [2].

Figure 4: Examples of generated images from the MHP Montage model and a naive approach. For each pair of images, the upper one is generated by a naive approach, and the lower one is by our proposed MHP Montage Model.

4.1.3 Evaluation Metrics. We follow [13] and use Average Precision based on Part (AP^p) and Percentage of Correctly Parsed Body Parts (PCP) are the evaluation metrics to evaluate the performance of multi-human parsing. For instance segmentation of closely entangled persons, we use the forward and backward scores [12] as the metric.

4.2 Experimental Analysis

4.2.1 Ablation Study for MHPM. In this subsection, we perform ablation study of the proposed MHPM to analyze different components within it.

MHP Montage Model. In MHPM, the MHP Montage model aims to provide additional training data for multi-human parsing. Here we explore its contribution to the overall multi-human parsing performance. We use a fixed setting of the MHP Solver, and only vary the training data in this experiment. We first explore if the MHP Montage model can generate a useful multi-human parsing dataset from existing single-human parsing datasets. We use scene images from the filtered background pool mentioned above. We use person images from the ATR dataset [17] to provide the foreground image pool. Totally, $30k$ images are generated and we term this dataset as Synthetic-Dataset 0 (SD0). Then we use the foreground images of persons in the training set of the MHP dataset to synthesize new multi-human images to aid the learning of multi-human parsing algorithms. Here we compare with a naive way of image composition, where the foreground images are randomly scaled and placed onto the background scene images, and we term this dataset as Synthetic Dataset 1 (SD1). Finally the synthetic dataset from our proposed MHP Montage model is termed Synthetic Dataset 2 (SD2). In both SD1 and SD2, there are about $10k$ synthetic images. We also generate more images to further increase the size of training data in SD2 to form a larger dataset ($SD2^+$) with about $30k$ images. The model trained with the original MHP training set is also used as a baseline comparison. We report the performance on the validation

Figure 5: Visualization of multi-human parsing results from the MHP Solver and other models. We show (a) the original image, (b) the ground truth global human parsing map, (c) the predicted global human parsing map from the MHP Solver, the person instance map from (d) MH-Parser [13], (e) Mask-RCNN [8], (f) Discriminative Loss [5], and (g) our proposed MHP Solver. On the global human parsing maps, each color represents a semantic category detailed by the legend. On the person instance map, each color stands for one person instance.

set of the MHP dataset, and the performance of the MHP Solver trained on different training datasets is listed in Tab. 1.

Table 1: Performance of the MHP Solver trained on different training datasets. The performance on the validation set of the MHP dataset is reported.

MHP Solver	$AP^p_{0.5}$	AP^p_{vol}	$PCP_{0.5}$
SD0	13.31	25.71	19.20
MHP train	53.05	50.12	52.99
MHP train + SD1	55.01	51.39	54.62
MHP train + SD2	56.19	51.12	55.17
MHP train + $SD2^+$	59.39	52.42	57.48

From the table, we can see that without a multi-human parsing dataset, a multi-human parsing model can still be trained with data generated by our MHP Montage model. However, the performance is worse compared with a model trained by a human annotated multi-human parsing dataset. Nevertheless, a fully annotated multi-human dataset is still necessary as we need such a dataset to benchmark the performance of multi-human parsing models. Compared to the baseline of training with only the training set of the MHP dataset, the synthetic dataset from the MHP Montage model plays a positive role in improving the performance, and the synthetic data from our MHP Montage model outperforms the naive approach of data synthesis.

We visually illustrate some of the images in SD1 and SD2 in Fig. 4. We can see that compared to the naively generated images, the MHP Montage model can adjust the scales and locations of each person in the background scene, so that the resultant images look more realistic and are more beneficial to training multi-human parsing algorithms.

MHP Solver. We also perform ablation analysis on the MHP Solver. In the MHP Solver, there are two groups of push and pull losses, *i.e.* group level and individual level. Each group can actually work on its own. We perform additional experiments which only use the group level losses or individual level losses. The performance is reported in Tab. 3.

We can see from the table that the GIPP loss in our MHP Solver outperforms both the group level loss and the individual level loss. By aggregating the forces in both the group level loss and the individual level loss, the proposed GIPP builds synergy from them and achieves performance gain in terms of all the metrics.

We report the AP^p and PCP of the MHP Solver under different IOU thresholds in Fig. 6. From the figure, we can see that both metrics drop monotonically as IOU threshold increases. Especially,

Table 2: Results from different methods on the MHP test set. The results of other models are obtained from the baseline methods in [13]. All denotes the entire test set, and Top 20% and Top 5% denote two challenging subsets of testing images with top 20% and top 5% largest overlaps between person instances, respectively.

	All			Top 20%			Top 5%		
	$AP^p_{0.5}$	AP^p_{vol}	$PCP_{0.5}$	$AP^p_{0.5}$	AP^p_{vol}	$PCP_{0.5}$	$AP^p_{0.5}$	AP^p_{vol}	$PCP_{0.5}$
Mask RCNN [8]	52.68	49.81	51.87	31.49	40.16	37.31	24.25	35.63	28.77
DL [5]	47.76	47.73	49.21	34.81	44.06	40.59	29.52	43.52	33.70
MH-Parser [13]	50.10	48.96	50.70	41.67	46.70	44.74	33.69	46.57	37.01
MHP Solver	51.07	49.30	52.65	37.28	44.40	43.21	30.79	43.71	35.08
MHPM	56.45	51.60	56.26	40.82	46.20	45.40	41.95	46.87	43.98

Table 3: Performance of the MHP Solver with different groups of losses. The performance on the validation set of the MHP dataset is reported.

	$AP^p_{0.5}$	AP^p_{vol}	$PCP_{0.5}$
MHP Solver w. G-only	51.87	50.06	52.06
MHP Solver w. I-only	50.97	49.13	51.12
MHP Solver w. GIPP	53.05	50.12	52.99

the values of AP^p and PCP are very low when the IOU threshold is high, *e.g.* IOU = 0.8. It demonstrates although the performance of multi-human parsing is good at low precision regime, it is still challenging to perform multi-human parsing at high precision levels.

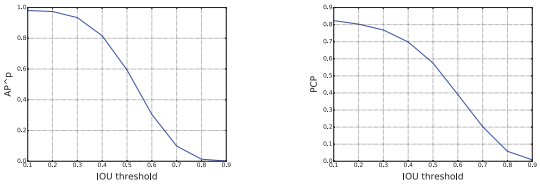

Figure 6: The AP^p and PCP under various IOU thresholds.

4.2.2 *Comparison with State-of-the-Arts.*

The MHP Dataset. In this subsection, we compare the results from our MHPM with other state-of-the-art methods in the literature. We consider Mask-RCNN [8], Discriminative Loss (DL) [5] and MH-Parser [13] as our baseline models. The Mask-RCNN is a person instance segmentation method, which only outputs instance segmentation results. Therefore the person instance segmentation masks from Mask-RCNN are aggregated to form person instance maps, which are combined with external global human parsing maps to generate multi-human parsing results. For DL and MH-Parser, they are both bottom-up methods which learn embeddings for pixels and superpixels, respectively, and use clustering to find person instances. The performance in terms of AP^p and PCP of different algorithms is listed in Tab. 2. Note that our MHP Solver uses the same trunk network ResNet-101, which is the same as all the other methods listed in the table.

From the results, we can see that our MHP Solver achieves the state-of-the-art performance on the multi-human parsing task. As a bottom-up method to distinguish person instances, our method outperforms Mask-RCNN, which is a top-down method, especially when the persons are closely entangled as in the case of Top 20% and Top 5% in Tab. 2. Our MHPM also outperforms other bottom-up methods like DL and MH-Parser, due to the carefully designed GIPP loss and the additional data obtained effortlessly from the MHP Montage model.

We visualize some multi-human parsing results in Fig. 5. We can see that our MHP Solver can capture the fine details of different person instances and assign pixels to the correct person instance, especially at the boundaries between persons.

The Buffy Dataset. Our MHP Solver can differentiate closely entangled persons, and we demonstrate this on the Buffy dataset as in [12]. Following [13] and [12], we evaluate the forward and backward score of Episode 4, 5 and 6 in the Buffy dataset and report the average values over the three episodes. Our MHPM is not trained on this dataset and only evaluation is performed on it. The results are in Tab. 4. We can see that our MHPM achieves the best performance on the Buffy dataset in separating closely entangled person instances.

Table 4: Performance of the MHP Solver on the Buffy dataset. We report the forward and backward scores from our MHP Solver and other state-of-the-art methods in the literature.

	Forward Score	Backward Score
Jiang *et.al.* [12]	68.22	69.66
MH-Parser [13]	71.11	71.94
MHP Solver	75.12	73.02

5 CONCLUSION

In this work, we proposed the Multi-Human Parsing Machines (MHPM) system, a one-stop solution to the Multi-Human Parsing (MHP) problem. The MHPM contains an MHP Montage model which intelligently composes human parsing images onto scene images to automatically generate new training data for multi-human parsing, and an MHP Solver which addresses the difficulties of distinguishing closely entangled person instances. With the proposed MHPM, we improve the state-of-the-art performances on the task of multi-human parsing.

ACKNOWLEDGMENTS

The work of Jianshu Li was partially funded by National Research Foundation of Singapore. The work of Jiashi Feng was partially supported by NUS IDS R-263-000-C67-646, ECRA R-263-000-C87-133 and MOE Tier-II R-263-000-D17-112.

REFERENCES

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

[2] 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:1606.00915* (2016).

[3] 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.

[4] 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.

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

[6] 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).

[7] 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.

[8] 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.

[9] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 770–778.

[10] 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).

[11] Max Jaderberg, Karen Simonyan, Andrew Zisserman, et al. 2015. Spatial transformer networks. In *Advances in neural information processing systems*. 2017–2025.

[12] 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.

[13] Jianshu Li, Jian Zhao, Yunchao Wei, Congyan Lang, Yidong Li, and Jiashi Feng. 2017. Towards Real World Human Parsing: Multiple-Human Parsing in the Wild. *arXiv preprint arXiv:1705.07206* (2017).

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

[15] 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.

[16] 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.

[17] 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.

[18] Chen-Hsuan Lin, Ersin Yumer, Oliver Wang, Eli Shechtman, and Simon Lucey. 2018. ST-GAN: Spatial Transformer Generative Adversarial Networks for Image Compositing. *arXiv preprint arXiv:1803.01837* (2018).

[19] 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.

[20] Alejandro Newell, Zhiao Huang, and Jia Deng. 2017. Associative embedding: End-to-end learning for joint detection and grouping. In *Advances in Neural Information Processing Systems*. 2274–2284.

[21] 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

[22] Yi-Hsuan Tsai, Xiaohui Shen, Zhe Lin, Kalyan Sunkavalli, Xin Lu, and Ming-Hsuan Yang. 2017. Deep image harmonization. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[23] 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.

[24] Huikai Wu, Shuai Zheng, Junge Zhang, and Kaiqi Huang. 2017. Gp-gan: Towards realistic high-resolution image blending. *arXiv preprint arXiv:1703.07195* (2017).

[25] 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.

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

[27] Jian Zhao, Jianshu Li, Xuecheng Nie, Fang Zhao, Yunpeng Chen, Zhecan Wang, Jiashi Feng, and Shuicheng Yan. 2017. Self-supervised neural aggregation networks for human parsing. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*. 7–15.

[28] Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. 2017. Places: A 10 million Image Database for Scene Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2017).

[29] Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. 2016. Semantic understanding of scenes through the ade20k dataset. *arXiv preprint arXiv:1608.05442* (2016).

[30] Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. 2017. Scene Parsing through ADE20K Dataset. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.