Understanding Humans in Crowded Scenes: Deep Nested Adversarial Learning and A New Benchmark for Multi-Human Parsing

Jian Zhao[1,2*] Jianshu Li[1*] Yu Cheng[1*] Li Zhou[1] Terence Sim[1] Shuicheng Yan[1,3] Jiashi Feng[1]
[1]National University of Singapore [2]National University of Defense Technology [3]Qihoo 360 AI Institute
{zhaojian90, jianshu}@u.nus.edu chengyu996@gmail.com zhouli2025@gmail.com
tsim@comp.nus.edu.sg {eleyans, elefjia}@nus.edu.sg

Abstract

Despite the noticeable progress in perceptual tasks like detection, instance segmentation and human parsing, computers still perform unsatisfactorily on visually understanding humans in crowded scenes, such as group behavior analysis, person re-identification and autonomous driving, etc. *To this end, models need to comprehensively perceive the semantic information and the differences between instances in a multi-human image, which is recently defined as the* multi-human parsing *task. In this paper, we present a new large-scale database "**M**ulti-**H**uman **P**arsing (MHP)" for algorithm development and evaluation, and advances the state-of-the-art in understanding humans in crowded scenes. MHP contains 25,403 elaborately annotated images with 58 fine-grained semantic category labels, involving 2-26 persons per image and captured in real-world scenes from various viewpoints, poses, occlusion, interactions and background. We further propose a novel deep **N**ested **A**dversarial **N**etwork (NAN) model for multi-human parsing. NAN consists of three **G**enerative **A**dversarial **N**etwork (GAN)-like sub-nets, respectively performing semantic saliency prediction, instance-agnostic parsing and instance-aware clustering. These sub-nets form a nested structure and are carefully designed to learn jointly in an end-to-end way. NAN consistently outperforms existing state-of-the-art solutions on our MHP and several other datasets, and serves as a strong baseline to drive the future research for multi-human parsing.*

1. Introduction

One of the primary goals of intelligent human-computer interaction is understanding the humans in visual scenes. It involves several perceptual tasks including detection, *i.e.* localizing different persons at a coarse, bounding box level (Fig. 1 (a)), instance segmentation, *i.e.* labelling each pixel of each person uniquely (Fig. 1 (b)), and human parsing, *i.e.* decomposing persons into their semantic categories (Fig. 1 (c)). Recently, deep learning based methods have achieved remarkable sucess in these perceptual tasks thanks to the availability of plentiful annotated images for training and evaluation purposes [10, 11, 30, 17].

Though exciting, current progress is still far from the

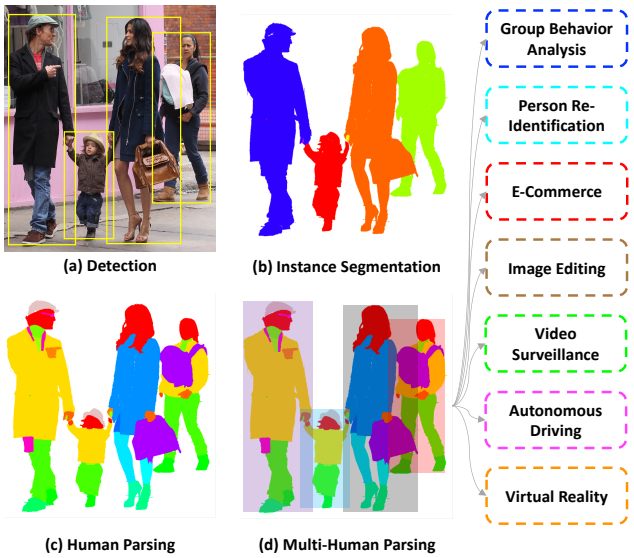


Figure 1: Illustration of motivation. While existing efforts on human-centric analysis have been devoted to (a) detection (localizing different persons at a coarse, bounding box level), (b) instance segmentation (labelling each pixel of each person uniquely) or (c) human parsing (decomposing persons into their semantic categories), we focus on (d) multi-human parsing (parsing body parts and fashion items at the instance level), which aligns better with many real-world applications. We introduce a new lage-scale, richly-annotated **M**ulti-**H**uman **P**arsing (MHP) dataset consisting of images with various viewpoints, poses, occlusion, human interactions and background. We further propose a novel deep **N**ested **A**dversarial **N**etwork (NAN) model for solving the challenging multi-human parsing problem effectively and efficiently. Best viewed in color.

*indicades equal contributions.

Table 1: Statistics for publicly available human parsing datasets.

Datasets	Instance Aware?	# Total	# Training	# Validation	# Testing	# Category
Buffy [39]	✓	748	452	-	296	13
Fashionista [43]	✗	685	456	-	229	56
PASCAL-Person-Part [4]	✗	3,533	1,716	-	1,817	7
ATR [28]	✗	17,700	16,000	700	1,000	18
LIP [17]	✗	50,462	30,462	10,000	10,000	20
MHP v1.0 [25]	✓	4,980	3,000	1,000	980	19
MHP v2.0	✓	25,403	15,403	5,000	5,000	59

utimate goal of visually understanding humans. As Fig. 1 shows, previous efforts on understanding humans in visual scenes either only consider coarse information or are agnostic to different instances. In the real-world scenarios, it is more likely that there simutaneously exist multiple persons, with various human interactions, poses and occlusion. Thus, it is more practically demanded to parse human body parts and fashion items at the instance level, which is recently defined as the *multi-human parsing* task [25]. Multi-human parsing enables more detailed understanding of humans in crowded scenes and aligns better with many real-world applications, such as group behavior analysis [15], person re-identification [47], e-commerce [38], image editing [42], video surveillance [6], autonomous driving [7] and virtual reality [29]. However, the existing benchmark datasets [10, 11, 30, 17] are not suitable for such a new task. Even though Li *et al.* [25] proposed a preliminary **M**ulti-**H**uman **P**arsing (MHP v1.0) dataset, it only contains 4,980 images annotated with 18 semantic labels. In this work, we propose a new large-scale benchmark "**M**ulti-**H**uman **P**arsing (MHP v2.0)[1]", aiming to push the frontiers of multi-human parsing research towards holistically understanding humans in crowded scenes. The data in MHP v2.0 cover wide variability and complexity *w.r.t.* viewpoints, poses, occlusion, human interactions and background. It in total includes 25,403 human images with pixel-wise annotations of 58 semantic categories.

We further propose a novel deep **N**ested **A**dversarial **N**etwork (NAN) model for solving the challenging multi-human parsing problem. Unlike most existing methods [25, 21, 26] which rely on separate stages of instance localization, human parsing and result refinement, the proposed NAN parses semantic categories and differentiates different person instances simultaneously in an effective and time-efficient manner. NAN consists of three **G**enerative **A**dversarial **N**etwork (GAN)-like sub-nets, respectively performing semantic saliency prediction, instance-agnostic parsing and instance-aware clustering. Each sub-task is simpler than the original multi-human parsing task, and is more easily addressed by the corresponding sub-net. Unlike many multi-task learning applications, in our method the sub-nets depend on each other, forming a causal nest by dynamically boosting each other through an adversarial strategy (See Fig. 5), which is hence called a "nested adversarial learning" structure. Such a structure enables effortless gradient **B**ackpro**P**agation (BP) in NAN such that it can be trained in a holistic, end-to-end way, which is favorable to both accuracy and speed. We conduct qualitative and quantitative experiments on the MHP v2.0 dataset proposed in this work, as well as the MHP v1.0 [25], PASCAL-Person-Part [4] and Buffy [39] benchmark datasets. The results demonstrate the superiority of NAN on multi-human parsing over the state-of-the-arts.

Our contributions are summarized as follows.

- We propose a new large-scale benchmark and evaluation server to advance understanding of humans in crowded scenes, which contains 25,403 images annotated pixel-wisely with 58 semantic category labels.
- We propose a novel deep **N**ested **A**dversarial **N**etwork (NAN) model for multi-human parsing, which serves as a strong baseline to inspire more future research efforts on this task.
- Comprehensive evaluations on the MHP v2.0 dataset proposed in this work, as well as the MHP v1.0 [25], PASCAL-Person-Part [4] and Buffy [39] benchmark datasets verify the superiority of NAN on understanding humans in crowded scenes over the state-of-the-arts.

2. Related Work

Human Parsing Datasets The statistics of popular publicly available datasets for human parsing are summarized in Tab. 1. The Buffy [39] dataset was released in 2011 for human parsing and instance segmentation. It contains only 748 images annotated with 13 semantic categories. The Fashionista [43] dataset was released in 2012 for human parsing, containing limited images annotated with 56 fashion categories. The PASCAL-Person-Part [4] dataset was initially annotated by Chen *et al.* [4] from the PASCAL-VOC-2010 [12] dataset. Chen *et al.* [3] extended it for human parsing with 7 coarse body part labels. The ATR [28] dataset was released in 2015 for human parsing with a large number of images annotated with 18 semantic categories. The LIP [17] dataset further extended ATR [28] by cropping person instances from Microsoft COCO [30]. It is a large-scale human parsing dataset with densely pixel-wise annotations of 20 semantic categories. But it has two

[1]The dataset is available at http://lv-mhp.github.io/

Figure 2: Annotation examples for our "**M**ulti-**H**uman **P**arsing (MHP v2.0)" dataset and existing datasets. (a) Examples in LIP [17]. LIP is restricted to an instance-agnostic setting and has limited semantic category annotations. (b) Examples in MHP v1.0 [25]. MHP v1.0 has lower scalability, variability and complexity, and only contains coarse labels. (c) Examples in our MHP v2.0. MHP v2.0 contains fine-grained semantic category labels with various viewpoints, poses, occlusion, interactions and background, aligned better with reality. Best viewed in color.

limitations. 1) Despite the large data size, it contains limited semantic category annotations, which restricts the fine-grained understanding of humans in visual scenes. 2) In LIP [17], only a small proportion of images involve multiple persons with interactions. Such an instance-agnostic setting severely deviates from reality. Even in the MHP v1.0 dataset proposed by Li *et al.* [25] for multi-human parsing, only 4,980 images are included and annotated with 18 semantic labels. Comparatively, our MHP v2.0 dataset contains 25,403 elaborately annotated images with 58 fine-grained semantic part labels. It is the largest and most comprehensive multi-human parsing dataset to date, to our best knowledge. Visual comparisons between LIP [17], MHP v1.0 [25] and our MHP v2.0 are provided in Fig. 2.

Human Parsing Approaches Recently, many research efforts have been devoted to human parsing [27, 17, 31, 46, 19, 9, 25, 21, 26] due to its wide range of potential applications. For example, Liang *et al.* [27] proposed a proposal-free network for instance segmentation by directly predicting the instance numbers of different categories and the pixel-level information. Gong *et al.* [17] proposed a self-supervised structure-sensitive learning approach, which imposes human pose structures to parsing results without resorting to extra supervision. Liu *et al.* [31] proposed a single frame video parsing method which integrates frame parsing, optical flow estimation and temporal fusion into a unified network. Zhao *et al.* [46] proposed a self-supervised neural aggregation network, which learns to aggregate the multi-scale features and incorporates a self-supervised joint loss to ensure the consistency between parsing and pose. He *et al.* [19] proposed the Mask R-CNN, which is extended from Faster R-CNN [34] by adding a branch for predicting an object mask in parallel with the existing branch for bounding box recognition. Brabandere *et al.* [9] proposed to tackle instance segmentation with a discriminative loss function, operating at the pixel level, which encourages a convolutional network to produce a representation of the image that can be easily clustered into instances with a simple post-processing step. However, these methods either only consider coarse semantic information or are agnostic to different instances. To enable more detailed human-centric analysis, Li *et al.* [25] initially proposed the multi-human parsing task, which aligns better with the realistic scenarios. They also proposed a novel MH-Parser model as a reference method which generates parsing maps and instance masks simutaneously in a bottom-up fashion. Jiang *et al.* [21] proposed a new approach to segment human instances and label their body parts using region assembly. Li *et al.* [26] proposed a framework with a human detector and a category-level segmentation module to segment the parts of objects at the instance level. These methods involve mutiple separate stages for instance localization, human parsing and

result refinement. In comparison, the proposed NAN produces accurate multi-human parsing results through a single forward-pass in a time-efficient manner without tedious pre- or post-processing.

3. Multi-Human Parsing Benchmark

In this section, we introduce the "**M**ulti-**H**uman **P**arsing (MHP v2.0)", a new large-scale dataset focusing on semantic understanding of humans in crowded scenes with several appealing properties. 1) It contains 25,403 elaborately annotated images with 58 fine-grained labels on body parts, fashion items and one background label, which is larger and more comprehensive than previous similar attempts [39, 25]. 2) The images within MHP v2.0 are collected from real-world scenarios, involving humans with various viewpoints, poses, occlusion, interactions and resolution. 3) The background of images in MHP v2.0 is more complex and diverse than previous datasets. Some examples are showed in Fig. 2. The MHP v2.0 dataset is expected to provide a new benchmark suitable for multi-human parsing together with a standard evaluation server where the test set will be kept secret to avoid overfitting.

3.1. Image Collection and Annotation

We manually specify some underlying relationships (such as family, couple, team, *etc.*) and possible scenes (such as sports, conferences, banquets, *etc.*) to ensure the diversity of returned results. Based on any one of these specifications, corresponding multi-human images are located by performing Internet searches over Creative Commons licensed imagery. For each identified image, the contained human number and the corresponding URL are stored in a spreadsheet. Automated scrapping software is used to download the multi-human imagery and stores all relevant information in a relational database. Moreover, a pool of images containing clearly visible persons with interactions and rich fashion items is also constructed from the existing human-centric datasets [44, 5, 45, 36, 22][2] to augment and complement Internet scraping results.

After curating the imagery, manual annotation is conducted by professional data annotators, which includes two distinct tasks. The first task is manually counting the number of foreground persons and duplicating each image to several copies according to the count number. Each duplicated image is marked with the image ID, the contained person number and a self-index. The second is assigning the fine-grained pixel-wise label to each semantic category for each person instance. We implement an annotation tool and generate multi-scale superpixels of images based on [2] to speed up the annotation. See Fig. 4 for an example. Each multi-human image contains at least two instances. The annotation for each instance is done in a left-to-right order, corresponding to the duplicated image with the self-index from beginning to end. For each instance, 58 semantic categories are defined and annotated, including *cap/hat*, *helmet*, *face*, *hair*, *left-arm*, *right-arm*, *left-hand*, *right-hand*, *protector*, *bikini/bra*, *jacket/windbreaker/hoodie*, *t-shirt*, *polo-shirt*, *sweater*, *singlet*, *torso-skin*, *pants*, *shorts/swim-shorts*, *skirt*, *stockings*, *socks*, *left-boot*, *right-boot*, *left-shoe*, *right-shoe*, *left-highheel*, *right-highheel*, *left-sandal*, *right-sandal*, *left-leg*, *right-leg*, *left-foot*, *right-foot*, *coat*, *dress*, *robe*, *jumpsuits*, *other-full-body-clothes*, *headwear*, *backpack*, *ball*, *bats*, *belt*, *bottle*, *carrybag*, *cases*, *sunglasses*, *eyewear*, *gloves*, *scarf*, *umbrella*, *wallet/purse*, *watch*, *wristband*, *tie*, *other-accessaries*, *other-upper-body-clothes* and *other-lower-body-clothes*. Each instance has a complete set of annotations whenever the corresponding category appears in the current image. When annotating one instance, others are regarded as background. Thus, the resulting annotation set for each image consists of N instance-level parsing masks, where N is the number of persons in the image.

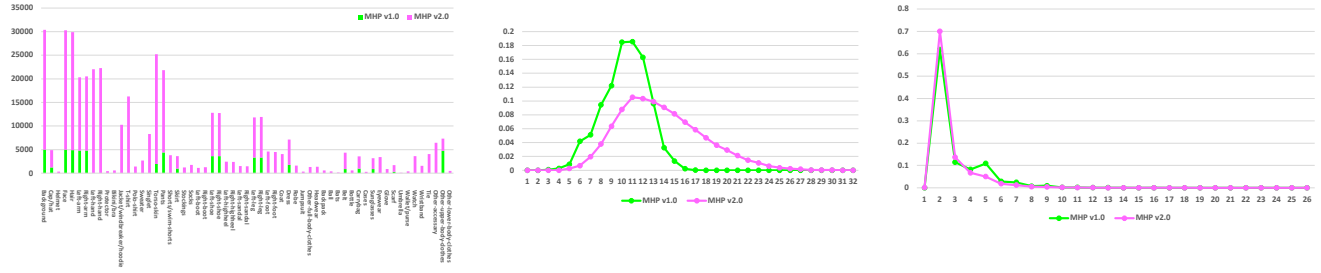
After annotation, manual inspection is performed on all images and corresponding annotations to verify the correctness. In cases where annotations are erroneous, the information is manually rectified by 5 well informed analysts. The whole work took around three months to accomplish by 25 professional data annotators.

3.2. Dataset Splits and Statistics

In total, there are 25,403 images in the MHP v2.0 dataset. Each image contains 2-26 person instances, with 3 on average. The resolution of the images ranges from 85×100 to 4,511×6,919, with 644×718 on average. We spit the images into training, validation and testing sets. Following random selection, we arrive at a unique split consisting of 15,403 training and 5,000 validation images with publicly available annotations, as well as 5,000 testing images with annotations withheld for benchmarking purpose.

The statistics *w.r.t.* data distribution on 59 semantic categories, the average semantic category number per image and the average instance number per image in the MHP v2.0 dataset are illustrated in Fig. 3 (a), (b) and (c), respectively. In general, *face*, arms and legs are the most remarkable parts of a human body. However, understanding humans in crowded scenes needs to analyze fine-grained details of each person of interest, including different body parts, clothes and accessaries. We therefore define 11 body parts, and 47 clothes and accessaries. Among these 11 body parts, we divide arms, hands, legs and feet into left and right side for more precise analysis, which also increases the difficulty of the task. We define *hair*, *face* and *torso-skin* as the remaining three body parts, which can be used

[2]PASCAL-VOC-2012 [11] and Microsoft COCO [30] are not included due to limited percent of crowd-scene images with fine details of persons.



(a) The data distribution on semantic category labels of our MHP v2.0 and the MHP v1.0 datasets.

(b) The statistic comparison on the average semantic category number per image between our MHP v2.0 and the MHP v1.0 datasets.

(c) The statistic comparison on the average instance number per image between our MHP v2.0 and the MHP v1.0 datasets.

Figure 3: Dataset statistics. Best viewed in color.

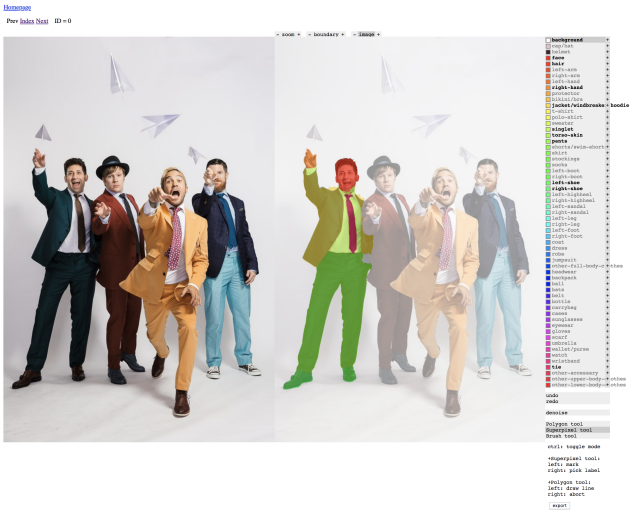


Figure 4: Annotation tool for multi-human parsing. Best viewed in color.

as auxiliary guidance for more comprehensive instance-level analysis. As for clothing categories, we have common clothes like *coat*, *jacket/windbreaker/hoodie*, *sweater*, *singlet*, *pants*, *shorts/swim-shorts* and shoes, confusing categories such as *t-shirt v.s. polo-shirt*, *stockings v.s. socks*, *skirt v.s. dress* and *robe*, and boots *v.s.* sandals and high-heels, and infrequent categories such as *cap/hat*, *helmet*, *protector*, *bikini/bra*, *jumpsuits*, *gloves* and *scarf*. Furthermore, accessaries like *sunglasses*, *belt*, *tie*, *watch* and bags are also taken into account, which are common but hard to predict, especially for the small-scale ones.

To summarize, the pre-defined semantic categories of MHP v2.0 involve most body parts, clothes and accessaries of different styles for men, women and children in all seasons. The images in the MHP v2.0 dataset contain diverse instance numbers, viewpoints, poses, occlusion, interactions and background complexities. MHP v2.0 aligns better with real-world scenarios and serves as a more realistic benchmark for human-centric analysis, which pushes the frontiers of fine-grained multi-human parsing research.

4. Deep Nested Adversarial Networks

As shown in Fig. 5, the proposed deep **N**ested **A**dversarial **N**etwork (NAN) model consists of three GAN-like sub-nets that jointly perform semantic saliency prediction, instance-agnostic parsing and instance-aware clustering end-to-end. NAN produces accurate multi-human parsing results through a single forward-pass in a time-efficient manner without tedious pre- or post-processing. We now present each component in details.

4.1. Semantic Saliency Prediction

Large modality and interaction variations are the main challenge to multi-human parsing and also the key obstacle to learning a well-performing human-centric analysis model. To address this problem, we propose to decompose the original task into three granularities and adaptively impose a prior on the specific process, each with the aid of a GAN-based sub-net. This reduces the training complexity and leads to better empirical performance with limited data.

The first sub-net estimates semantic saliency maps to locate the most noticeable and eye-attracting human regions in images, which serves as a basic prior to facilitate further processing on humans, as illustrated in Fig. 5 left. We formulate semantic saliency prediction as a binary pixel-wise labelling problem to segment out foreground *v.s.* background. Inspired by the recent success of **F**ully **C**onvolutional **N**etwork**s** (FCNs) [32] based methods on image-to-image applications [24, 19], we leverage an FCN backbone (FCN-8s [32]) as the generator $\mathrm{G1}_{\theta_1} : \mathbb{R}^{H\times W\times C} \mapsto \mathbb{R}^{H\times W\times C'}$ of NAN for semantic saliency prediction, where θ_1 denotes the network parameters, and H, W, C and C' denote the image height, width, channel number and semantic category (*i.e.*, foreground plus background) number, repectively.

Formally, let the input RGB image be denoted by x and the semantic saliency map be denoted by x', then

$$x' := \mathrm{G1}_{\theta_1}(x). \tag{1}$$

The key requirements for G1 are that the semantic saliency map x' should present indistinguishable properi-

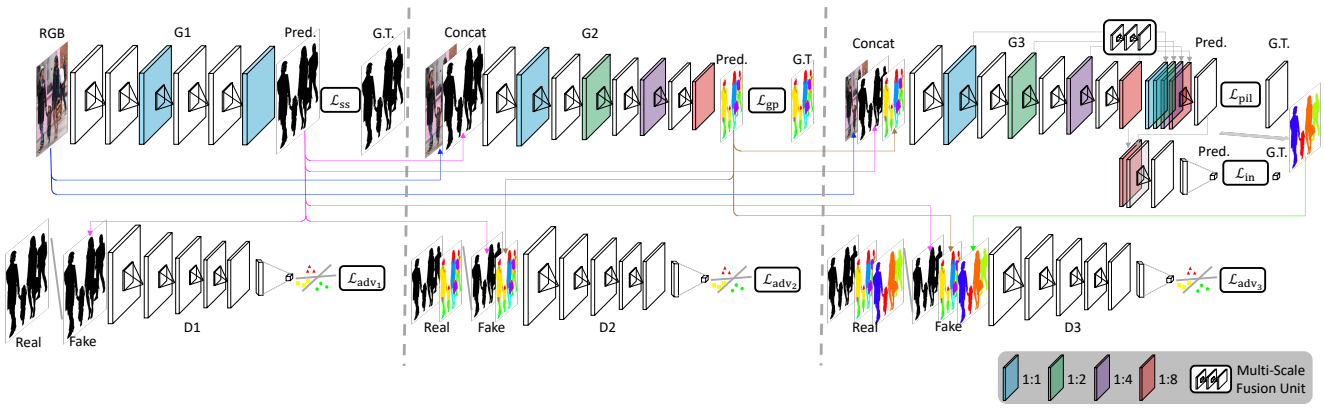


Figure 5: Deep Nested Adversarial Networks (NAN) for multi-human parsing. NAN consists of three GAN-like sub-nets, respectively performing semantic saliency prediction, instance-agnostic parsing and instance-aware clustering. Each sub-task is simpler than the original multi-human parsing task, and is more easily addressed by the corresponding sub-net. The sub-nets depend on each other, forming a causal nest by dynamically boosting each other via an adversarial strategy. Such a structure enables effortless gradient Back**P**ropagation (BP) of NAN such that it can be trained in a holistic, end-to-end way. NAN produces accurate multi-human parsing results through a single forward-pass in a time-efficient manner without tedious pre- or post-processing. Best viewed in color.

ties compared with a real one (*i.e.*, ground truth) in appearance while preserving the intrinsic contextually remarkable information.

To this end, we propose to learn θ_1 by minimizing a combination of two losses:

$$\mathcal{L}_{\mathrm{G1}_{\theta_1}} = -\lambda_1 \mathcal{L}_{\mathrm{adv}_1} + \lambda_2 \mathcal{L}_{\mathrm{ss}}, \tag{2}$$

where $\mathcal{L}_{\mathrm{adv}_1}$ is the **adv**ersarial loss for refining realism and alleviating artifacts, $\mathcal{L}_{\mathrm{ss}}$ is the **s**emantic **s**aliency loss for pixel-wise image labelling, λ are weighting parameters among different losses.

$\mathcal{L}_{\mathrm{ss}}$ is a pixel-wise cross-entropy loss calculated based on the binary pixel-wise annotations to learn θ_1:

$$\mathcal{L}_{\mathrm{ss}} = \mathcal{L}_{\mathrm{ss}}(X'(\theta_1)|X). \tag{3}$$

$\mathcal{L}_{\mathrm{adv}_1}$ is proposed to narrow the gap between the distributions of generated and real results. To facilitate this process, we leverage a **C**onvolutional **N**eural **N**etwork (CNN) backbone as the discriminator $\mathrm{D1}_{\phi_1} : \mathbb{R}^{H\times W\times C'} \mapsto \mathbb{R}^1$ to be as simple as possible to avoid typical GAN tricks. We alternatively optimize $\mathrm{G1}_{\theta_1}$ and $\mathrm{D1}_{\phi_1}$ to learn θ_1 and ϕ_1:

$$\begin{cases} \mathcal{L}^{\mathrm{G1}}_{\mathrm{adv}_1} = \mathcal{L}_{\mathrm{adv}_1}(K(\theta_1)|X'(\theta_1), X'_{\mathrm{GT}}), \\ \mathcal{L}^{\mathrm{D1}}_{\mathrm{adv}_1} = \mathcal{L}_{\mathrm{adv}_1}(K(\phi_1)|X'(\theta_1), X'_{\mathrm{GT}}), \end{cases} \tag{4}$$

where K denotes the binary real *v.s.* fake indicator.

4.2. Instance-Agnostic Parsing

The second sub-net concatenates the information from the original RGB image with semantic saliency prior as input and estimates a fine-grained instance-agnostic parsing map, which further serves as stronger semantic guidance from the global perspective to facilitate instance-aware clustering, as illustrated in Fig. 5 middle. We formulate instance-agnostic parsing as a multi-class dense classification problem to mask semantically consistent regions of body parts and fashion items. Inspired by the leading performance of the skip-net on recognition tasks [40, 20], we modify a skip-net (WS-ResNet [40]) into an FCN-based architecture as the generator $\mathrm{G}_{2\theta_2} : \mathbb{R}^{H\times W\times (C+C')} \mapsto \mathbb{R}^{H/8\times W/8\times C''}$ of NAN to learn a highly non-linear transformation for instance-agnostic parsing, where θ_2 denotes the network parameters for the generator and C'' denotes the semantic category number. The prediction is downsampled by 8 for accuracy *v.s.* speed trade-off. Contextual information from global and local regions compensates each other and naturally benefits human parsing. The hierarchical features within a skip-net are multi-scale in nature due to the increasing receptive field sizes, which are combined together via skip connections. Such a combined representation comprehensively maintains the contextual information, which is crucial for generating smooth and accurate parsing results.

Formally, let the instance-agnostic parsing map be denoted by x'', then

$$x'' := \mathrm{G2}_{\theta_2}(x \cup x'). \tag{5}$$

Similar to the first sub-net, we propose to learn θ_2 by minimizing:

$$\mathcal{L}_{\mathrm{G2}_{\theta_2}} = -\lambda_3 \mathcal{L}_{\mathrm{adv}_2} + \lambda_4 \mathcal{L}_{\mathrm{gp}}, \tag{6}$$

where $\mathcal{L}_{\mathrm{gp}}$ is the **g**lobal **p**arsing loss for semantic part labelling.

$\mathcal{L}_{\mathrm{gp}}$ is a standard pixel-wise cross-entropy loss calculated based on the multi-class pixel-wise annotations to

learn θ_2. θ_1 is also slightly finetuned due to the hinged gradient backpropagation route within the nested structure:

$$\mathcal{L}_{\text{gp}} = \mathcal{L}_{\text{gp}}(X''(\theta_2, \theta_1)|X \cup X'(\theta_1)). \tag{7}$$

$\mathcal{L}_{\text{adv}_2}$ is proposed to ensure the correctness and realism of the current phase and also the previous one for information flow consistency. To facilitate this process, we leverage a same CNN backbone with D1_{ϕ_1} as the discriminator $\text{D2}_{\phi_2} : \mathbb{R}^{H\times W\times(C'+C'')} \mapsto \mathbb{R}^1$, which are learned separately. We alternatively optimize G2_{θ_2} and D2_{ϕ_2} to learn θ_2, ϕ_2 and slightly finetune θ_1:

$$\begin{cases} \mathcal{L}^{\text{G2}}_{\text{adv}_2} = \mathcal{L}_{\text{adv}_2}(K(\theta_2)|X'(\theta_1) \cup X''(\theta_2, \theta_1), X'_{\text{GT}} \cup X''_{\text{GT}}), \\ \mathcal{L}^{\text{D2}}_{\text{adv}_2} = \mathcal{L}_{\text{adv}_2}(K(\phi_2)|X'(\theta_1) \cup X''(\theta_2, \theta_1), X'_{\text{GT}} \cup X''_{\text{GT}}). \end{cases} \tag{8}$$

4.3. Instance-Aware Clustering

The third sub-net concatenates the information from the original RGB image with semantic saliency and instance-agnostic parsing priors as input and estimates an instance-aware clustering map by associating each semantic parsing mask to one of the person instances in the scene, as illustrated in Fig. 5 right. Inspired by the observation that a human glances at an image and instantly knows how many and where the objects are in the image, we formulate instance-aware clustering by parallelly inferring the instance number and pixel-wise instance location, discarding the requirement of time-consuming region proposal generation. We modify a same backbone architecture G2_{θ_2} to incorporate two sibling branches as the generator $\text{G}_{3\theta_3} : \mathbb{R}^{H/8\times W/8\times(C+C'+C'')} \mapsto \mathbb{R}^{H/8\times W/8\times M} \cup \mathbb{R}^1$ of NAN for location-sensitive learning, where θ_3 denotes the network parameters for the generator and M denotes the pre-defined instance location coordinate number. As multi-scale features integrating both global and local contextual information are crucial for increasing location prediction accuracy, we further augment the pixel-wise instance location prediction branch with a **M**ulti-**S**cale **F**usion **U**nit (MSFU) to fuse shallow-, middle- and deep-level features, while using the feature maps downsampled by 8 concatenated with feature maps from the first branch for instance number regression.

Formally, let the pixel-wise instance location map be denoted by $\tilde{x}$ and the instance number be denoted by n, then

$$\tilde{x} \cup n := \text{G3}_{\theta_3}(x \cup x' \cup x''). \tag{9}$$

We propose to learn θ_3 by minimizing:

$$\mathcal{L}_{\text{G3}_{\theta_3}} = -\lambda_5\mathcal{L}_{\text{adv}_3} + \lambda_6\mathcal{L}_{\text{pil}} + \lambda_7\mathcal{L}_{\text{in}}, \tag{10}$$

where $\mathcal{L}_{\text{pil}}$ is the **p**ixel-wise **i**nstance **l**ocation loss for pixel-wise instance location regression and $\mathcal{L}_{\text{in}}$ is the **i**nstance **n**umber loss for instance number regression.

$\mathcal{L}_{\text{pil}}$ is a standard smooth-ℓ_1 loss [16] calculated based on the foreground pixel-wise instance location annotations to learn θ_3. Since a person instance can be identified by its top-left corner (x^l, y^l) and bottom-right corner (x^r, y^r) of the surrounding bounding box, for each pixel belonging to the person instance, the pixel-wise instance location vector is defined as $[x^l/w, y^l/h, x^r/w, y^r/h]$, where w and h are the width and height of the person instance for normalization, respectively. $\mathcal{L}_{\text{in}}$ is a standard ℓ_2 loss calculated based on the instance number annotations to learn θ_3. θ_2 and θ_1 are also slightly finetuned due to the chained schema within the nest:

$$\begin{cases} \mathcal{L}_{\text{pil}} = \mathcal{L}_{\text{pil}}(\tilde{X}(\theta_3, \theta_2, \theta_1)|X \cup X'(\theta_1) \cup X''(\theta_2, \theta_1)), \\ \mathcal{L}_{\text{in}} = \mathcal{L}_{\text{in}}(N(\theta_3, \theta_2, \theta_1)|X \cup X'(\theta_1) \cup X''(\theta_2, \theta_1)). \end{cases} \tag{11}$$

Given these information, instance-aware clustering maps can be effortlessly generated with little computational overhead, which are denoted by $\hat{X} \in \mathbb{R}^{M'}$. Similar to $\mathcal{L}_{\text{adv}_2}$, $\mathcal{L}_{\text{adv}_3}$ is proposed to ensure the correctness and realism of all phases for the information flow consistency. To facilitate this process, we leverage a same CNN backbone with D2_{ϕ_2} as the discriminator $\text{D3}_{\phi_3} : \mathbb{R}^{H\times W\times(C'+C''+M')} \mapsto \mathbb{R}^1$, which are learned separately. We alternatively optimize G3_{θ_3} and D3_{ϕ_3} to learn θ_3, ϕ_3 and slightly finetune θ_2 and θ_1:

$$\begin{cases} \mathcal{L}^{\text{G3}}_{\text{adv}_3} = \mathcal{L}_{\text{adv}_3}(K(\theta_3)|X'(\theta_1) \cup X''(\theta_2, \theta_1) \cup \hat{X}(\theta_3, \theta_2, \theta_1), X'_{\text{GT}} \cup X''_{\text{GT}} \cup \hat{X}_{\text{GT}}), \\ \mathcal{L}^{\text{D3}}_{\text{adv}_3} = \mathcal{L}_{\text{adv}_3}(K(\phi_3)|X'(\theta_1) \cup X''(\theta_2, \theta_1) \cup \hat{X}(\theta_3, \theta_2, \theta_1), X'_{\text{GT}} \cup X''_{\text{GT}} \cup \hat{X}_{\text{GT}}). \end{cases} \tag{12}$$

4.4. Training and Inference

The goal of NAN is to use sets of real targets to learn three GAN-like sub-nets that mutually boost and jointly accomplish multi-human parsing. Each separate loss serves as a deep supervision within the nested structure benefitting network convergence. The overall objective function for NAN is

$$\mathcal{L}_{\text{NAN}} = -\sum_{i=1}^{3}\lambda_i\mathcal{L}_{\text{adv}_i} + \lambda_4\mathcal{L}_{\text{ss}} + \lambda_5\mathcal{L}_{\text{gp}} + \lambda_6\mathcal{L}_{\text{pil}} + \lambda_7\mathcal{L}_{\text{in}}. \tag{13}$$

Clearly, the NAN is end-to-end trainable and can be optimized with the proposed nested adversarial learning strategy and BP algorithm.

During testing, we simply feed the input image X into NAN to get the instance-agnostic parsing map X'' from G2_{θ_2}, pixel-wise instance location map $\tilde{X}$ and instance number N from G3_{θ_3}. Then we employ an off-the-shelf clustering method [33] to obtain the instance-aware clustering map $\hat{X}$. Example results are visualized in Fig. 6.

5. Experiments

We evaluate NAN qualitatively and quantitatively under various settings and granularities for understanding humans in crowded scenes. In particular, we evaluate multi-

human parsing performance on the MHP v2.0 dataset proposed in this work, as well as the MHP v1.0 [25] and PASCAL-Person-Part [4] benchmark datasets. We also evaluate instance-agnostic parsing and instance-aware clustering results on the Buffy [39] benchmark dataset, which are byproducts of NAN.

5.1. Experimental Settings

5.1.1 Implementation Details

Throughout the experiments, the sizes of the RGB image X, the semantic saliency prediction X', inputs to the discriminator $\mathrm{D1}_{\phi_1}$ and inputs to the generator $\mathrm{G2}_{\theta_2}$ are fixed as 512×512; the sizes of the instance-agnostic parsing prediction X'', instance-aware clustering prediction $\hat{X}$, inputs to the discriminator $\mathrm{D2}_{\phi_2}$, inputs to the generator $\mathrm{G3}_{\theta_3}$, inputs to the discriminator $\mathrm{D3}_{\phi_3}$ and instance location map $\tilde{X}$ are fixed as 64×64; the channel number of the pixel-wise instance location map is fixed as 4, incorporating two corner points of the associated bounding box; the constraint factors $\lambda_i, i \in \{1,2,3,4,5,6,7\}$ are empirically fixed as $0.01, 0.01, 0.01, 1.00, 1.00, 10.00$ and 1.00, respectively; the generator $\mathrm{G1}_{\theta_1}$ is initialized with FCN-8s [32] by replacing the last layer with a new convolutional layer with kernel size $1 \times 1 \times 2$, pretrained on PASCAL-VOC-2011 [13] and finetuned on the target dataset; the generator $\mathrm{G2}_{\theta_2}$ is initialized with WS-ResNet [40] by eliminating the spatial pooling layers, increasing the strides of the first convolutional layers up to 2 in $\mathrm{B}_i, i \in \{2,3,4\}$, eliminating the top-most global pooling layer and the linear classifier, and adding two new convolutional layers with kernel sizes $3 \times 3 \times 512$ and $1 \times 1 \times C''$, pretrained on ImageNet [35] and PASCAL-VOC-2012 [11], and finetuned on the target dataset; the generator $\mathrm{G3}_{\theta_3}$ is initialized with the same backbone architecture and pre-trained weights with $\mathrm{G2}_{\theta_2}$ (which are learned separately), by further augmenting it with two sibling branches for pixel-wise instance location map prediction and instance number prediction, where the first branch utilizes a MSFU (three convolutional layers with kernal sizes $3 \times 3 \times i, i \in \{128, 128, 4\}$ for specific scale adaption) ended with a convolutional layer with kernel size $1 \times 1 \times 4$ for multi-scale feature aggregation and a final convolutional layer with kernel size $1 \times 1 \times 4$ for location regression and the second branch utilizes the feature maps downsampled by 8 concatenated with the feature maps from the first branch ended with a global pooling layer, a hidden 512-way fully-connected layer and a final 1-way fully-connected layer for instance number regression; the three discriminators $\mathrm{Di}_{\phi_i}, i \in \{1,2,3\}$ (which are learned separately) are all initialized with a VGG-16 [37] by adding a new convolutional layer at the very begining with kernel size $1 \times 1 \times 3$ for input adaption, and replacing the last layer with a new 1-way fully-connected layer activated by sigmoid, pre-trained on ImageNet [35] and finetuned on the target dataset; the newly added layers are randomly initialized by drawing weights from a zero-mean Gaussian distribution with standard deviation 0.01; we employ an off-the-shelf clustering method [33] to obtain the instance-aware clustering map $\hat{X}$; the dropout ratio is empirically fixed as 0.7; the weight decay and batch size are fixed as 5×10^{-3} and 4, respectively; We use an initial learning rate of 1×10^{-6} for pre-trained layers, and 1×10^{-4} for newly added layers in all our experiments; we decrease the learning rate to $1/10$ of the previous one after 20 epochs and train the network for roughly 60 epochs one after the other; the proposed network is implemented based on the publicly available TensorFlow [1] platform, which is trained using Adam (β_1=0.5) on four NVIDIA GeForce GTX TITAN X GPUs with 12G memory; the same training setting is utilized for all our compared network variants; we evaluate the testing time by averaging the running time for images on the target set on NVIDIA GeForce GTX TITAN X GPU and Intel Core i7-4930K CPU@3.40GHZ; our NAN can rapidly process one 512×512 image in about 1 second, which compares much favorably to other state-of-the-art approaches, as the current state-of-the-art methods [25, 21, 26] rely on region proposal preprocessing and complex processing steps.

5.1.2 Evaluation Metrics

Following [25], we use the **A**verage **P**recision based on **p**art (AP^p) and **P**ercentage of **C**orrectly parsed semantic **P**arts (PCP) metrics for multi-human parsing evaluation. Different from the **A**verage **P**recision based on **r**egion (AP^r) used in instance segmentation [27, 18], AP^p uses part-level pixel **I**ntersection **o**ver **U**nion (IoU) of different semantic part categories within a person instance to determine if one instance is a true positive. We prefer AP^p over AP^r as we focus on human-centric analysis and we aim to investigate to how well a person instance as a whole is parsed. Additionally, we also report the AP^p_{vol}, which is the mean of the AP^p at IoU thresholds ranging from 0.1 to 0.9, in increments of 0.1. As AP^p averages the IoU of each semantic part category, it fails to reflect how many semantic parts are correctly parsed. We further incorporate the PCP, originally used in human pose estimation [14, 4], to evaluate the parsing quality within person instances. For each true-positive person instance, we find all the semantic categories (excluding background) with pixel IoU larger than a threshold, which are regarded as correctly parsed. The PCP of one person instance is the ratio between the correctly parsed semantic category number and the total semantic category number of that person. Missed person instances are assigned with 0 PCP. The overall PCP is the average PCP for all person instances. Note that PCP is also a human-centric evaluation metric.

Table 2: Component analysis on the MHP v2.0 validation set.

Setting	Method	$AP^p_{0.5}(\%)$	$AP^p_{vol}(\%)$	$PCP_{0.5}(\%)$
Baseline	Mask R-CNN [19]	14.50	33.51	25.12
	MH-Parser [25]	18.05	35.87	26.91
Network Structure	w/o G1	22.67	38.11	31.95
	G2 w/o concatenated input	21.88	36.79	29.02
	G3 w/o concatenated input	22.36	35.92	25.48
	w/o D1	23.81	33.95	27.59
	w/o D2	19.02	29.66	22.89
	D2 w/o concatenated input	21.55	31.94	24.90
	w/o D3	20.62	32.83	26.22
	D3 w/o concatenated input	21.80	34.54	27.30
	w/o MSFU	18.76	26.62	24.94
Ours	NAN	24.83	42.77	34.37
Upperbound	X'_{GT}	26.17	43.59	38.11
	X''_{GT}	28.98	48.55	38.03
	N_{GT}	28.39	47.76	39.25
	$\tilde{X}_{\mathrm{GT}}$	30.18	51.44	41.18

5.2. Evaluations on the MHP v2.0 Benchmark

The MHP v2.0 dataset proposed in this paper is the largest and most comprehensive multi-human parsing benchmark to date, which extends MHP v1.0 [25] to push the frontiers of understanding humans in crowded scenes by containing 25,403 elaborately annotated images with 58 fine-grained semantic category labels. Annotation examples are visualized in Fig. 2 (c). The data are randomly organized into 3 splits, consisting of 15,403 training and 5,000 validation images with publicly available annotations, as well as 5,000 testing images with annotations withheld for benchmarking purpose. Evaluation systems report the AP^p and PCP over the validation and testing sets.

5.2.1 Component Analysis

We first investigate different architectures and loss function combinations of NAN to see their respective roles in multi-human parsing. We compare 16 variants from four aspects, *i.e.*, different baselines (Mask R-CNN [19][3] and MH-Parser [25]), different network structures (w/o G1, G2 w/o concatenated input (RGB only), G3 w/o concatenated input (RGB only), w/o D1, w/o D2, D2 w/o concatenated input, w/o D3, D3 w/o concatenated input, w/o MSFU), our proposed NAN, and upperbounds (X'_{GT}: use the ground truth semantic saliency maps instead of G1 prediction while keeping other settings the same; X''_{GT}: use the ground truth instance-agnostic parsing maps instead of G2 prediction while keeping other settings the same; N_{GT}: use the ground truth instance number instead of G3 prediction while keeping other settings the same; $\tilde{X}_{\mathrm{GT}}$: use the ground truth pixel-wise instance location maps instead of G3 prediction while keeping other settings the same).

The performance comparison in terms of AP^p@IoU=0.5, AP^p_{vol} and PCP@IoU=0.5 on the MHP v2.0 validation set is reported in Tab. 2. By comaring the results from the 1^{st} *v.s.* 3^{rd} panels, we observe that our proposed NAN consistently outperforms the baselines Mask R-CNN [19] and MH-Parser [25] by a large margin, *i.e.*, 10.33% and 6.78% in terms of AP^p, 9.26% and 6.90% in terms of AP^p_{vol}, and 9.25% and 7.46% in terms of PCP. Mask R-CNN [19] suffers difficulties to differentiate entangled humans. MH-Parser [25] involves multiple stages for instance localization, human parsing and result refinement with high complexity, yielding sub-optimal results, whereas NAN parses semantic categories, differentiates different person instances and refines results simultaneously through deep nested adversarial learning in an effective yet time-efficient manner. By comaring the results from the 2^{nd} *v.s.* 3^{rd} panels, we observe that NAN consistently outperforms the 9 variants in terms of network structure. In particular, w/o G1 refers to truncating the semantic saliency prediction sub-net from NAN, leading to 2.16%, 4.66% and 2.42% performance drop in terms of all metrics. This verifies the necessity of semantic saliency prediction that locates the most noticeable human regions in images to serve as a basic prior to facilitate further human-centic processing. The superiority of incorporating adaptive prior information to specific process can be verified by comparing $\mathrm{Gi}, i \in \{2, 3\}$ w/o concatenated input with NAN, *i.e.*, 2.95%, 5.98% and 5.35%; 2.47%, 6.85% and 8.89% differences in terms of all metrics. The superiority of incorporating adversarial learning to specific process can be verified by comparing w/o $\mathrm{Di}, i \in \{1, 2, 3\}$ with NAN, *i.e.*, 1.02%, 8.82% and 6.78%; 5.81%, 13.11% and 11.48%; 4.21%, 9.94% and 8.15% decrease in terms of all metrics. Nested adversarial learning strategy ensures the correctness and realism of all phases for information flow consistency, the superiority of which is verified by comparing $\mathrm{Di}, i \in \{2, 3\}$ w/o concatenated input with NAN, *i.e.*, 3.28%, 10.83% and 9.47%; 3.03%, 8.23% and 7.07% decline in terms of all metrics. MSFU dynamically fuses multi-scale features for enhancing instance-aware clustering accuracy, the superiority of which is verified by comparing w/o MSFU with NAN, *i.e.*, 6.07%, 16.15% and 9.43% drop in terms of all metrics. Finally, we also evaluate the limitations of our current algorithm. By comparing X'_{GT} with NAN, only 1.34%, 0.82% and 3.74% improvement in term of all metrics are obtained, which shows that the errors from semantic saliency prediction are already small and have only little effect on the final results. A large gap between 28.98%, 48.55% and 38.03% of X''_{GT} and 24.83%, 42.77% and 34.37% of NAN shows that a better instance-agnostic parsing network architecture can definitely help improve the performance of multi-human parsing under our NAN framework. By comparing N_{GT} and $\tilde{X}_{\mathrm{GT}}$ with NAN, 3.56%, 4.99% and 4.88%; 5.35%, 8.67% and 6.81% improvement in term of all metrics are obtained, which shows that accurate instance-aware clustering results are critical for superior multi-human parsing.

[3] As existing instance segmentation methods only offer silhouettes of different person instances, for comparison, we combine them with our instance-agnostic parsing prediction to generate the final multi-human parsing results.

5.2.2 Quantitative Comparison

The performance comparison of the proposed NAN with two state-of-the-art methods in terms of AP^p@IoU=0.5, AP^p_{vol} and PCP@IoU=0.5 on the MHP v2.0 testing set is reported in Tab. 3. Following [25], we conduct experiments under three settings: **All** reports the evaluation over the whole testing set; $\textbf{Inter}_{20\%}$ reports the evaluation over the sub-set containing the images with top 20% interaction intensity[4]; $\textbf{Inter}_{10\%}$ reports the evaluation over the sub-set containing the images with top 10% interaction intensity. Our NAN is significantly superior over other state-of-the-arts on setting-1. In particular, NAN improves the 2^{nd}-best by 7.15%, 5.70% and 5.27% in terms of all metrics. For the more challenging scenarios with intensive interactions (setting-2, 3), NAN also consistently achieves the best performance. In particular, for $\textbf{Inter}_{20\%}$ and $\textbf{Inter}_{10\%}$, NAN improves the 2^{nd}-best by 5.23%, 4.65% and 5.62%; 1.63%, 3.20% and 3.33% in terms of all metrics. This verifies the effectiveness of our NAN for multi-human parsing and understanding humans in crowded scenes. Moreover, NAN can rapidly process one 512×512 image in about 1 second with acceptable resource consumption, which is attractive to real applications. This compares much favorably to MH-Parser [25] (14.94 img/s), which relies on separate and complex post-processing (including CRF [23]) steps.

5.2.3 Qualitative Comparison

Fig. 6 visualizes the qualitative comparison of the proposed NAN with two state-of-the-art methods and corresponding ground truths on the MHP v2.0 dataset. Note that Mask R-CNN [19] only offers silhouettes of different person instances, we only compare our instance-aware clustering results with it while comparing our holistic results with MH-Parser [25]. It can be observed that the proposed NAN performs well in multi-human parsing with a wide range of viewpoints, poses, occlusion, interactions and background complexity. The instance-agnostic parsing and instance-aware clustering predictions of NAN present high consistency with corresponding ground truths, thanks to the novel network structure and effective training strategy. In contrast, Mask R-CNN [19] suffers difficulties to differentiate entangled humans, while MH-Parser [25] struggles to generate fine-grained parsing results and clearly segmented instance masks. This further desmonstrates the effectiveness of the proposed NAN. We also show some failure cases of our NAN in Fig. 7. As can be observed, humans in crowded scenes with heavy occlusion, extreme poses and intensive interactions are difficult to identify and segment. Some small-scale semantic categories within person instances are difficult to parse. This confirms that MHP v2.0 aligns with real-world situations and deserves more furture attention and research efforts.

[4]For each testing image, we calculate the pair-wise instance bounding box IoU and use the mean value as the interaction intensity for each image.

5.3. Evaluations on the MHP v1.0 Benchmark

The MHP v1.0[5] dataset is the first multi-human parsing benchmark, originally proposed by Li *et al.* [25], which contains 4,980 images annotated with 18 semantic labels. Annotation examples are visualized in Fig. 2 (b). The data are randomly organized into 3 splits, consisting of 3,000 training, 1,000 validation and 1,000 testing images with publicly available annotations. Evaluation systems report the AP^p and PCP over the testing set. Refer to [25] for more details.

The performance comparison of the proposed NAN with three state-of-the-art methods in terms of AP^p@IoU=0.5, AP^p_{vol} and PCP@IoU=0.5 on the MHP v1.0 [25] testing set is reported in Tab. 4. With the nested adversarial learning of semantic saliency prediction, instance-agnostic parsing and instance-aware clustering, our method outperforms the 2^{nd}-best by 4.41% for $\text{AP}^p_{0.5}$, 6.95% for AP^p_{vol} and 8.04% for $\text{PCP}_{0.5}$. Visual comparison of multi-human parsing results by NAN and three state-of-the-art methods is provided in Fig. 8, which further validates the advantages of our NAN over existing solutions.

5.4. Evaluations on the PASCAL-Person-Part Benchmark

The PASCAL-Person-Part[6] [4] dataset is a set of additional annotations for PASCAL-VOC-2010 [12]. It goes beyond the original PASCAL object detection task by providing pixel-wise labels for six human body parts, *i.e.*, *head*, *torso*, *upper-/lower-arms*, and *upper-/lower-legs*. The rest of each image is considered as background. There are 3,535 images in the PASCAL-Person-Part [4] dataset, which is split into separate training set containing 1,717 images and testing set containing 1,818 images. For fair comparison, we report the AP^r over the testing set for multi-human parsing. Refer to [3, 41] for more details.

The performance comparison of the proposed NAN with two state-of-the-art methods in terms of AP^r@IoU=$k_{k=0.5}^{0.7}$ and AP^r_{vol} on the PASCAL-Person-Part [4] testing set is reported in Tab. 5. Our method dramatically surpasses the 2^{nd}-best by 18.90% for $\text{AP}^r_{0.7}$ and 13.80% for AP^r_{vol}. Qualitative multi-human parsing results by NAN are visualized in Fig. 9, which possess a high concordance with corresponding ground truths. This again verifies the effectiveness of our method for human-centric analysis.

Table 3: Multi-human parsing quantitative comparison on the MHP v2.0 testing set.

Method	All			$\text{Inter}_{20\%}$			$\text{Inter}_{10\%}$			Speed (img/s)
	$\text{AP}^r_{0.5}(\%)$	$\text{AP}^r_{vol}(\%)$	$\text{PCP}_{0.5}(\%)$	$\text{AP}^r_{0.5}(\%)$	$\text{AP}^r_{vol}(\%)$	$\text{PCP}_{0.5}(\%)$	$\text{AP}^r_{0.5}(\%)$	$\text{AP}^r_{vol}(\%)$	$\text{PCP}_{0.5}(\%)$	
Mask R-CNN [19]	14.90	33.88	25.11	4.77	24.28	12.75	2.23	20.73	8.38	-
MH-Parser [25]	17.99	36.08	26.98	13.38	34.25	22.31	13.25	34.29	21.28	14.94
NAN	**25.14**	**41.78**	**32.25**	**18.61**	**38.90**	**27.93**	**14.88**	**37.49**	**24.61**	**1.08**

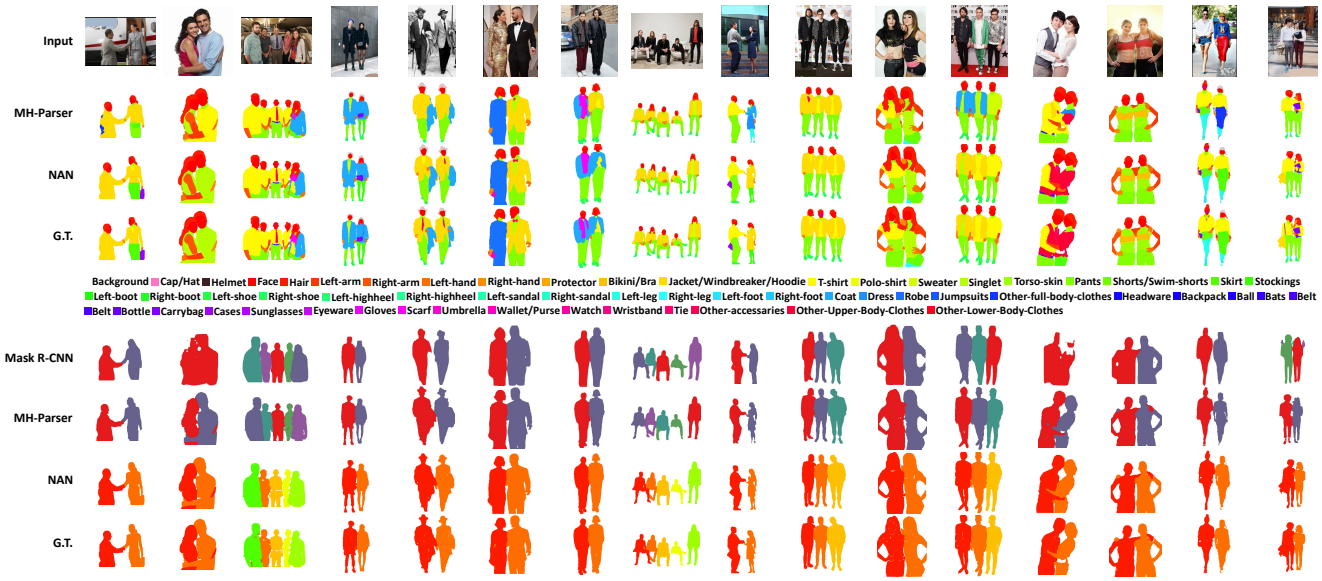


Figure 6: Multi-human parsing qualitative comparison on the MHP v2.0 dataset. Best viewed in color.

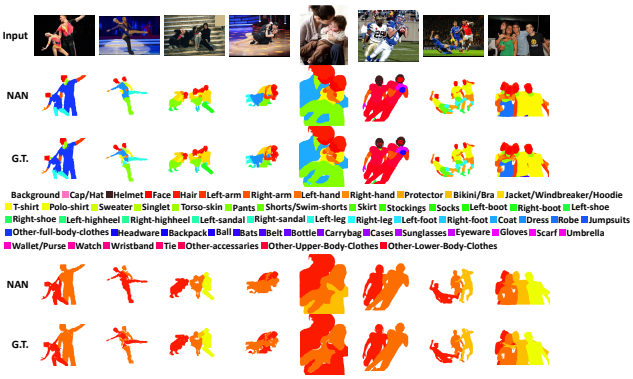


Figure 7: Failure cases of multi-human parsing results by our NAN on the proposed MHP v2.0 dataset. Best viewed in color.

Table 4: Multi-human parsing quantitative comparison on the MHP v1.0 [25] testing set.

Method	$\text{AP}^r_{0.5}(\%)$	$\text{AP}^r_{vol}(\%)$	$\text{PCP}_{0.5}(\%)$
DL [9]	47.76	47.73	49.21
MH-Parser [25]	50.10	48.96	50.70
Mask R-CNN [19]	52.68	49.81	51.87
NAN	**57.09**	**56.76**	**59.91**

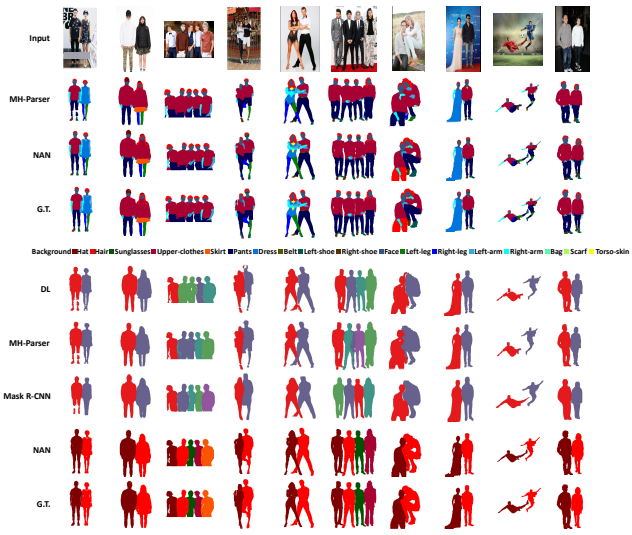


Figure 8: Multi-human parsing qualitative comparison on the MHP v1.0 [25] dataset. Best viewed in color.

5.5. Evaluations on the Buffy Benchmark

The Buffy[7] [39] dataset was released in 2011 for human parsing and instance segmentation, which contains 748 im-

[5]The dataset is available at http://lv-mhp.github.io/

[6]The dataset is available at http://www.stat.ucla.edu/~xianjie.chen/pascal_part_dataset/pascal_part.html

[7]The dataset is available at https://www.inf.ethz.ch/personal/ladickyl/Buffy.zip

Table 5: Multi-human parsing quantitative comparison on the PASCAL-Person-Part [4] testing set.

Method	$AP^r_{0.5}(\%)$	$AP^r_{0.6}(\%)$	$AP^r_{0.7}(\%)$	$AP^r_{vol}(\%)$
MNC [8]	38.80	28.10	19.30	36.70
Li *et al.* [26]	40.60	30.40	19.10	38.40
NAN	**59.70**	**51.40**	**38.00**	**52.20**

Figure 9: Multi-human parsing qualitative comparison on the PASCAL-Person-Part [4] dataset. Best viewed in color.

Table 6: Instance segmentation quantitative comparison on the Buffy [39] dataset episode 4, 5 and 6.

Method	F (%)	B (%)	Ave. (%)
Vineet *et al.* [39]	-	-	62.40
Jiang *et al.* [21]	68.22	69.66	68.94
MH-Parser [25]	71.11	71.94	71.53
NAN	**77.24**	**79.92**	**78.58**

Figure 10: Qualitative instance-agnostic parsing (upper panel) and instance-aware clustering (lower panel) results by NAN on the Buffy [39] dataset. Best viewed in color.

ages annotated with 12 semantic labels. The data are randomly organized into 2 splits, consisting of 452 training and 296 testing images with publicly available annotations. For fair comparison, we report the **F**orward (F) and **B**ackward (B) scores [21] over the episode 4, 5 and 6 for instance segmentation evaluation. Refer to [39, 21] for more details.

The performance comparison of the proposed NAN with three state-of-the-art methods in terms of F and B scores on the Buffy [39] dataset episode 4, 5 and 6 is reported in Tab. 6. Our NAN consistently achieves the best performance for all metrics. In particualr, NAN significantly improves the 2^{nd}-best by 6.13% for F score and 7.98% for B score, with an average boost of 7.05%. Qualitative instance-agnostic parsing and instance-aware clustering results by NAN are visualized in Fig. 10, which well shows the promising potential of our method for fine-grained understanding humans in crowded scenes.

6. Conclusions

In this work, we presented "**M**ulti-**H**uman **P**arsing (MHP v2.0)", a large-scale multi-human parsing dataset and a carefully designed benchmark to spark progress in understanding humans in crowded scenes. MHP v2.0 contains 25,403 images, which are richly labelled with 59 semantic categories. We also proposed a novel deep **N**ested **A**dversarial **N**etwork (NAN) model to address this challenging problem and performed detailed evaluations of the proposed method with current state-of-the-arts on MHP v2.0 and several other datasets. We envision the proposed MHP v2.0 dataset and the baseline method would drive the human parsing research towards real-world application scenario with simultaneous presence of multiple persons and complex interactions among them. In future, we will continue to take efforts to construct a more comprehensive multi-human parsing benchmark dataset with more images and more detailed semantic category annotations to further push the frontiers of multi-human parsing research.

Acknowledgement

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] M. Abadi, P. Barham, J. Chen, Z. Chen, A. Davis, J. Dean, M. Devin, S. Ghemawat, G. Irving, M. Isard, et al. Tensorflow: A system for large-scale machine learning. 8

[2] P. Arbelaez, M. Maire, C. Fowlkes, and J. Malik. Contour detection and hierarchical image segmentation. *TPAMI*, 33(5):898–916, 2011. 4

[3] L.-C. Chen, Y. Yang, J. Wang, W. Xu, and A. L. Yuille. Attention to scale: Scale-aware semantic image segmentation. In *CVPR*, pages 3640–3649, 2016. 2, 10

[4] X. Chen, R. Mottaghi, X. Liu, S. Fidler, R. Urtasun, and A. Yuille. Detect what you can: Detecting and representing objects using holistic models and body parts. In *CVPR*, pages 1971–1978, 2014. 2, 8, 10, 12

[5] X. Chu, W. Ouyang, W. Yang, and X. Wang. Multi-task recurrent neural network for immediacy prediction. In *ICCV*, pages 3352–3360, 2015. 4

[6] R. T. Collins, A. J. Lipton, T. Kanade, H. Fujiyoshi, D. Duggins, Y. Tsin, D. Tolliver, N. Enomoto, O. Hasegawa, P. Burt,

et al. A system for video surveillance and monitoring. *VSAM final report*, pages 1–68, 2000. 2

[7] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele. The cityscapes dataset for semantic urban scene understanding. In *CVPR*, pages 3213–3223, 2016. 2

[8] J. Dai, K. He, and J. Sun. Instance-aware semantic segmentation via multi-task network cascades. In *CVPR*, pages 3150–3158, 2016. 12

[9] B. De Brabandere, D. Neven, and L. Van Gool. Semantic instance segmentation with a discriminative loss function. *arXiv preprint arXiv:1708.02551*, 2017. 3, 11

[10] P. Dollar, C. Wojek, B. Schiele, and P. Perona. Pedestrian detection: An evaluation of the state of the art. *TPAMI*, 34(4):743–761, 2012. 1, 2

[11] M. Everingham, S. A. Eslami, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman. The pascal visual object classes challenge: A retrospective. *IJCV*, 111(1):98–136, 2015. 1, 2, 4, 8

[12] M. Everingham, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman. The pascal visual object classes (voc) challenge. *IJCV*, 88(2):303–338, 2010. 2, 10

[13] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The PASCAL Visual Object Classes Challenge 2011 (VOC2011) Results. http://www.pascal-network.org/challenges/VOC/voc2011/workshop/index.html. 8

[14] V. Ferrari, M. Marin-Jimenez, and A. Zisserman. Progressive search space reduction for human pose estimation. In *CVPR*, pages 1–8, 2008. 8

[15] C. Gan, M. Lin, Y. Yang, G. de Melo, and A. G. Hauptmann. Concepts not alone: Exploring pairwise relationships for zero-shot video activity recognition. In *AAAI*, page 3487, 2016. 2

[16] R. Girshick. Fast r-cnn. *arXiv preprint arXiv:1504.08083*, 2015. 7

[17] K. Gong, X. Liang, X. Shen, and L. Lin. Look into person: Self-supervised structure-sensitive learning and a new benchmark for human parsing. *arXiv preprint arXiv:1703.05446*, 2017. 1, 2, 3

[18] B. Hariharan, P. Arbeláez, R. Girshick, and J. Malik. Simultaneous detection and segmentation. In *ECCV*, pages 297–312. Springer, 2014. 8

[19] K. He, G. Gkioxari, P. Dollár, and R. Girshick. Mask r-cnn. In *ICCV*, pages 2980–2988, 2017. 3, 5, 9, 10, 11

[20] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016. 6

[21] H. Jiang and K. Grauman. Detangling people: Individuating multiple close people and their body parts via region assembly. *arXiv preprint arXiv:1604.03880*, 2016. 2, 3, 8, 12

[22] B. F. Klare, B. Klein, E. Taborsky, A. Blanton, J. Cheney, K. Allen, P. Grother, A. Mah, and A. K. Jain. Pushing the frontiers of unconstrained face detection and recognition: Iarpa janus benchmark a. In *CVPR*, pages 1931–1939, 2015. 4

[23] P. Krähenbühl and V. Koltun. Efficient inference in fully connected crfs with gaussian edge potentials. In J. Shawe-Taylor, R. S. Zemel, P. L. Bartlett, F. Pereira, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 24*, pages 109–117. Curran Associates, Inc., 2011. 10

[24] G. Li, Y. Xie, L. Lin, and Y. Yu. Instance-level salient object segmentation. In *CVPR*, pages 247–256, 2017. 5

[25] J. Li, J. Zhao, Y. Wei, C. Lang, Y. Li, T. Sim, S. Yan, and J. Feng. Multi-human parsing in the wild. *arXiv preprint arXiv:1705.07206*, 2017. 2, 3, 4, 8, 9, 10, 11, 12

[26] Q. Li, A. Arnab, and P. H. Torr. Holistic, instance-level human parsing. *arXiv preprint arXiv:1709.03612*, 2017. 2, 3, 8, 12

[27] X. Liang, Y. Wei, X. Shen, J. Yang, L. Lin, and S. Yan. Proposal-free network for instance-level object segmentation. *arXiv preprint arXiv:1509.02636*, 2015. 3, 8

[28] X. Liang, C. Xu, X. Shen, J. Yang, S. Liu, J. Tang, L. Lin, and S. Yan. Human parsing with contextualized convolutional neural network. In *ICCV*, pages 1386–1394, 2015. 2

[29] J. Lin, X. Guo, J. Shao, C. Jiang, Y. Zhu, and S.-C. Zhu. A virtual reality platform for dynamic human-scene interaction. In *SIGGRAPH*, page 11. ACM, 2016. 2

[30] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft coco: Common objects in context. In *ECCV*, pages 740–755, 2014. 1, 2, 4

[31] S. Liu, C. Wang, R. Qian, H. Yu, R. Bao, and Y. Sun. Surveillance video parsing with single frame supervision. In *CVPRW*, pages 1–9, 2017. 3

[32] J. Long, E. Shelhamer, and T. Darrell. Fully convolutional networks for semantic segmentation. In *CVPR*, pages 3431–3440, 2015. 5, 8

[33] A. Y. Ng, M. I. Jordan, and Y. Weiss. On spectral clustering: Analysis and an algorithm. In *NIPS*, pages 849–856, 2002. 7, 8

[34] S. Ren, K. He, R. Girshick, and J. Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *NIPS*, pages 91–99, 2015. 3

[35] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, et al. Imagenet large scale visual recognition challenge. *IJCV*, 115(3):211–252, 2015. 8

[36] B. Sapp and B. Taskar. Modec: Multimodal decomposable models for human pose estimation. In *CVPR*, pages 3674–3681, 2013. 4

[37] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 8

[38] E. Turban, D. King, J. Lee, and D. Viehland. Electronic commerce: A managerial perspective 2002. *Prentice Hall: ISBN 0*, 13(975285):4, 2002. 2

[39] V. Vineet, J. Warrell, L. Ladicky, and P. H. Torr. Human instance segmentation from video using detector-based conditional random fields. In *BMVC*, volume 2, pages 12–15, 2011. 2, 4, 8, 11, 12

[40] Z. Wu, C. Shen, and A. v. d. Hengel. Wider or deeper: Revisiting the resnet model for visual recognition. *arXiv preprint arXiv:1611.10080*, 2016. 6, 8

[41] F. Xia, P. Wang, L.-C. Chen, and A. L. Yuille. Zoom better to see clearer: Human and object parsing with hierarchical auto-zoom net. In *ECCV*, pages 648–663, 2016. 10

[42] N. Xu, B. Price, S. Cohen, J. Yang, and T. S. Huang. Deep interactive object selection. In *CVPR*, pages 373–381, 2016. 2

[43] K. Yamaguchi, M. H. Kiapour, L. E. Ortiz, and T. L. Berg. Parsing clothing in fashion photographs. In *CVPR*, pages 3570–3577, 2012. 2

[44] N. Zhang, M. Paluri, Y. Taigman, R. Fergus, and L. Bourdev. Beyond frontal faces: Improving person recognition using multiple cues. In *CVPR*, pages 4804–4813, 2015. 4

[45] Z. Zhang, P. Luo, C. C. Loy, and X. Tang. From facial expression recognition to interpersonal relation prediction. *IJCV*, pages 1–20, 2016. 4

[46] J. Zhao, J. Li, X. Nie, F. Zhao, Y. Chen, Z. Wang, J. Feng, and S. Yan. Self-supervised neural aggregation networks for human parsing. In *CVPRW*, pages 7–15, 2017. 3

[47] R. Zhao, W. Ouyang, and X. Wang. Unsupervised salience learning for person re-identification. In *CVPR*, pages 3586–3593, 2013. 2