Hierarchical Human Parsing with Explicit Part Relation Reasoning

Anonymous CVPR submission

Paper ID XXX

Abstract

We conduct extensive experiments on five popular human parsing datasets. Our parser achieves state-of-the-art results across the datasets. Our code will be made publicly available.

1. Introduction

Human parsing is to segment human bodies into semantic parts, *e.g.*, head, arm, leg, *etc*. It has attracted tremendous attention in the literature, as it enables fine-grained human understanding and benefits a wide spectrum of human-centric applications, such as human behavior analysis, human-robot interaction, and many others.

The rapid advances in deep learning and development of large-scale datasets [20, 50, 39, 28, 40, 19] have inspired a number of pioneering approaches in this filed. In general, further beyond intuitively employing existing well-designed semantic segmentation network architectures (*e.g.*, FCN [37], DeepLab [5], *etc*.) to this task [6, 58, 40], recent attempts typically leverage human structures as a high-level human configuration constraint [51, 16, 20, 50, 42, 59, 48]. Many such efforts resort to straightforwardly encoding human pose information into the parsing model [51, 16, 20, 50, 42], suffering from trivial structure information and extra human pose annotations. Some others consider top-down (coarse-to-fine) [59] or multi-level part [48] information over human structural layouts. Although claiming hierarchical or compositional relation aware, they do not consider distinct characteristics of different relation types. For example, the **compositional** and **decompositional** relations between constituent and entire parts (*e.g.*, {*upper body*, *lower body*} and *full body*), and the **dependency** relations between kinematically connected parts (*e.g.*, *hand* and *arm*) present significantly different relation rules.

In short summary, though above representative efforts greatly advance this field, three major limitations in human structures/relation modeling are still observed: **(1)** The utilized structure information is weak (in a form of sparse human keypoints) or the types of studied relations are incomplete (limited to compositional/decompositional relation). **(2)** The relation modeling strategies are over-general and simple; do not seem to characterize well the diverse body part relations. **(3)** Since the human body presents a tree-like cyclic topology, an iterative inference is desirable. However, current arts are built upon an immediate, feed-forward prediction scheme, largely ignoring this issue.

To response to above challenges and enable a deeper understating of human structures, we develop a new structured human parser that more precisely describes different kinds of part relations, and more efficiently reasons structures with the prism of a message-passing, feedback inference scheme. To address the first two issues, we start with in-depth and comprehensively analyzing three common relations, namely dependency, decomposition, and composition. Three distinct relation networks are elaborately designed and imposed to explicitly satisfy the corresponding dependency/decompositional/compositional constraints. Then, we construct our structured human parser as a tree-like, end-to-end trainable graph model, where the nodes represent the human parts, and edges are built upon our relational models for efficient relation reasoning. For the third issue, a modified, convolutional message passing procedure is performed over the human hierarchy, enabling our method have the chance to get better parsing results from a global view. All the components, *i.e.*, part nodes, edge (relational) functions, and message passing modules, are fully differentiable, enabling our whole framework end-to-end trainable and in turn facilitating learning about parts, relations, and inference algorithms.

More importantly, the suggested structured human parser can be viewed as an essential variant of message passing neural networks (MPNNs) [17, 46], yet significantly distinct in two aspects. (1) Most of previous MPNNs are edge-type-agnostic, while ours addresses relation-typed structure reasoning with better expressive capability. (2) By replacing Multilayer Perceptron (MLP) based MPNN units as convolutional neural network (CNN) based operations, our parser gains spatial information preserving property, which is desirable for such pixel-wise prediction task.

We extensively evaluate our approach using five standard human parsing datasets [20, 50, 39, 28, 40], achieving state-of-the-art performance on all of them (Sec.6.1). In addition, with ablation studies over essential components of our parser, three key insights are found: (1) Exploring different relations reside on human bodies are valuable for human parsing. (2) Distinctly and explicitly modeling different relation types would better support human structure reasoning. (3) Message passing based feed-back inference is able to reinforce parsing results.

2. Related Work

Human parsing: In the past decade, active research has been made on pixel-level human semantic understanding. Early approaches tend to leverage image regions [32, 53, 54], hand-crafted features [47, 8], part templates [2, 13, 12] and human keypoints [52, 32, 53, 54], and typically explore certain heuristics over human body configurations [3, 13, 12] in a CRF [52, 25], structured model [53, 13], grammar model [3, 12], or generative model [15, 44] framework.

Recent progress in this field is driven by the streamlined designs of deep learning architectures. Some pioneering efforts revisit classic template matching strategy [28, 33], address local and global cues [31], or use tree-LSTMs to gather structure information [29, 30]. However, due to the use of superpixel [31, 29, 30] or HOG feature [39], they are not truly end-to-end trainable and typically time-consuming. Consequent approaches thus follow an FCN architecture, which use multi-level cues [49], address feature aggregation [40, 58], employ adversarial learning [57, 41, 34], or leverage cross-domain knowledge [34, 18]. To further explore inherent structures, numerous approaches [51, 58, 20, 50, 16, 42] choose to encode pose information into the parser network, while relying on additional annotations. Some methods [59, 48, 18] instead fuse information over human structures. Though elegant, they largely ignore iterative inference [59, 48] and seldom address explicit relation modeling, easily suffering from weak expressive ability and risk of sub-optimal results.

With the general success of these works, we make a further step towards more precisely describing different relations reside on human bodies, *i.e.*, dependency, decomposition, and composition, and addressing iterative, spatial-information preserving inference over human hierarchy.

Graph neural networks: As a part of the huge graph learning family, graph neural networks (GNNs) have a rich history (dated back to [45]) and became a veritable explosion in research community in the last few years [21]. GNNs effectively learn graph representations in an end-to-end manner, which can generally be divided into two broad classes, called Graph Convolutional Networks (GCNs) and Message Passing Graph Networks (MPGNs), respectively. The former [14, 43, 24] seeks to directly extend classical CNNs to non-Euclidean data. The simple architecture facilitates its popularity, while limits its modeling capability of complex structures [21]. MPGNs [17, 46] parameterize all the nodes, edges, and all the information fusion steps in graph learning, leading to more complicated yet flexible architectures.

Our structured human parser, which falls in the second category, can be viewed as an early attempt that explores GNNs in human parsing area. In contrast to conventional MPGNs which are mainly MLP based and edge-type-agnostic, we provide a spatial information preserved and relation-type aware graph learning scheme.

3. Our Approach

3.1. Problem Definition

Formally, we represent the hierarchical human semantic structure as a directed graph $\mathcal{G}=(\mathcal{V},\mathcal{E},\mathcal{Y})$. Here the node set $\mathcal{V}=\mathcal{V}^1\cup\ldots\cup\mathcal{V}^L$ represents human parts in three different semantic levels, where the leaf nodes $\mathcal{V}^1$ are the most fine-grained parts (such as *head, arm, hand, etc.*), typically considered in common human parsers, $\mathcal{V}^2=\{$*upper-body, lower-body*$\}$ two middle-level nodes, and $\mathcal{V}^3=\{$*full-body*$\}$ the root node. The edge set $\mathcal{E}\in\binom{\mathcal{V}}{2}$ represents the relations among human parts (nodes), *i.e.*, the directed edge $e=(u,v)\in\mathcal{E}$ links node u to $v: u\rightarrow v$. Each node v is associated with a feature vector (node embedding) $\boldsymbol{h}_v$, and each edge (u,v) also has a corresponding edge embedding $\boldsymbol{h}_{u,v}$. The label attached to node v is a part groundtruth map $y_v\in\mathcal{Y}$, which is expected to be predicted by $\boldsymbol{h}_v$.

Our human parser is trained in a graph learning scheme. With the full supervision (the groundtruth map $\mathcal{Y}$) from exiting human parsing datasets, our parser learns the node representations $\{\boldsymbol{h}_v\}_{v\in\mathcal{V}}$ by reasoning structures at the level of individual parts and their relations, and iteratively fusing information over $\mathcal{G}$. It is worth mentioning that the use of the hierarchical groundtruth representation does not introduce any extra annotation requirement, as high-level node annotations can be simply obtained by combining the lower-level labels according to the compositional relations.

3.2. Structured Human Paring Network

Node embedding: As an initial step, a learnable project function is used to map the input image representation into node (part) features, in order to obtain sufficient expressive power. More specifically, let us denote the input image feature as $\boldsymbol{x}\in\mathbb{R}^{H\times W\times C}$, which is from a DeepLabV3-like backbone network, and the project function as $P\colon\mathbb{R}^{H\times W\times C}\mapsto\mathbb{R}_{\geq 0}^{H\times W\times c\times|\mathcal{V}|}$, where $|\mathcal{V}|$ indicates the number of nodes. The node embeddings $\{\boldsymbol{h}_v\in\mathbb{R}_{\geq 0}^{H\times W\times c}\}_{v\in\mathcal{V}}$ are initialized by:

$$\{\boldsymbol{h}_v\}_{v\in\mathcal{V}}=P(\boldsymbol{x}),\tag{1}$$

where each node embedding $\boldsymbol{h}_v$ is a (H, W, c)-dimensional tenor that encodes full spatial details.

Next we focus on analyzing three kinds of human part relations, *i.e.*, dependency, decomposition, and composition, and elaborating our parameterized relation modeling.

Parameterized human part relation modeling: Basically, an edge embedding $\boldsymbol{h}_{u,v}$ captures the relations between nodes u and v. In a high level, most of representative GNNs [23, 17] and previous structured human parsers [59, 48] work in an edge-type-agnostic manner, *i.e.*, a unified, shared relation network $R : \mathbb{R}_{\geq 0}^{H\times W\times c} \times \mathbb{R}_{\geq 0}^{H\times W\times c} \mapsto \mathbb{R}_{\geq 0}^{H\times W\times c}$ is used to capture all the relations: $\boldsymbol{h}_{u,v} = R(\boldsymbol{h}_u, \boldsymbol{h}_v)$. Such strategy may lose discriminability of different relation types and is too general to address geometric and anatomical constrains in such specific task. In contrast, we formulate $\boldsymbol{h}_{u,v}$ as a relation-typed manner:

$$\boldsymbol{h}_{u,v} = R^r(F^r(\boldsymbol{h}_u), \boldsymbol{h}_v), \tag{2}$$

where $r \in \{\text{dep}, \text{dec}, \text{com}\}$. $F^r(\cdot)$ is a set of attention-based relation-adaption operations, which is used to enhance the original node represent $\boldsymbol{h}_u$ by addressing specific relation constraints. From a view of information diffusion mechanism in graph theory [45], for the edge (u, v) that links a starting node u to an ending node v, it indicates v receives incoming information (*i.e.*, $\boldsymbol{h}_{u,v}$) from u. Thus we use $F^r(\cdot)$ to make $\boldsymbol{h}_u$ better accommodate the target v. R^r is edge-type specific, with the more tractable feature $F^r(\boldsymbol{h}_u)$ of the related part u, more expressive relation features for v are expected to be easily obtained and further benefit final parsing results. In this way, we can learn more sophisticated and impressive relation patterns within human bodies.

Dependency relation modeling: Dependency relation is defined as pairwise, kinematic relations between human parts, such as {*head*, *torso*}, {*upper-leg*, *lower-leg*}. In the human hierarchy $\mathcal{G}$, dependency relations are represented as those horizontal edges connecting sibling nodes.

Dependency relation has two key characteristics in image domain: the regions of kinematically connected parts: u and v, are **(1)** spatially adjacent or very close; and **(2)** mutually exclusive, *i.e.*, a pixel can only be recognized as either u or v. For node u, with its kinematically connected siblings $\mathcal{K}_u$, a dependency relation network R^{dep} is designed as:

$$\begin{aligned} \boldsymbol{h}_{u,v} &= R^{\text{dep}}(F^{\text{dep}}(\boldsymbol{h}_u), \boldsymbol{h}_v), \quad v \in \mathcal{K}_u, \\ F^{dep}(\boldsymbol{h}_u) &= F^{\text{cont}}(\boldsymbol{h}_u) \odot (\mathbf{1} - A(\boldsymbol{h}_u)), \end{aligned} \tag{3}$$

where $F^{\text{cont}}(\boldsymbol{h}_u) \in \mathbb{R}_{\geq 0}^{H\times W\times c}$ is a *contextual information extraction* module that captures local structures around u for addressing the first characteristic, $(\mathbf{1} - A(\boldsymbol{h}_u)) \in [0, 1]^{H\times W\times c}$ is an *exclusive attention* that addresses the second characteristic, and '$\odot$' indicates the attention-based feature enhancement operation.

F^{cont} is formed by a multilayered deformable convolutional network [10] with $XX \times XX$ convolution kernel and a large dilation values. The rationale behind this choice are three-fold. First, from an intuitive view, as u and v are spatially close, we can adapt a sliding window strategy over $\boldsymbol{h}_u$ for 'collecting' local contextual information that are useful for v. This is analogous to applying a convolutional network that convolves every position in $\boldsymbol{h}_u$, where the receptive field can be viewed as the local searching window. Deformable convolutional network is favored as it allows a learnable receptive field, *i.e.*, automatically find a suitable searching window. Second, by adaptively placing the convolution kernel, deformable convolutional network shows stronger structure modeling capacity. Third, deformable convolutional network is able to achieve a large receptive field (local searching window) with a small amount of parameters.

$A(\boldsymbol{h}_u) : \mathbb{R}_{\geq 0}^{H\times W\times C} \mapsto [0, 1]^{H\times W}$ produces an activation map of $\boldsymbol{h}_u$, by computing statistics of $\boldsymbol{h}_u$ across the channel dimension:

$$A(\boldsymbol{h}_u) = \frac{1}{Z}\sum\nolimits_{d=1}^{D} (\boldsymbol{h}_u^d)^\ell, \tag{4}$$

where $\boldsymbol{h}_u^d$ denotes a slice of the $\boldsymbol{h}_u$ in d-th channel, and $Z = \texttt{max}(\sum_{d=1}^{D}(\boldsymbol{h}_u^d)^\ell)$ is a normalization factor. $A(\boldsymbol{h}_u)$ is a non-parametric, activation-based attention [55], without bringing any extra training effort. $\mathbf{1} - A(\boldsymbol{h}_u)$ acts as an attention mechanism that addresses the mutually exclusive property by suppressing the response within the regions belonging to u. We empirically set $\ell = 2$ and do not observe significant performance change with $\ell = \{1, 2, 3\}$.

Decompositional relation modeling: Decompositional relations are represented by those vertical edges starting from parent nodes to corresponding child nodes in the human hierarchy $\mathcal{G}$, such as a parent node *full-body* can be separated into {*upper-body*, *lower-body*}, and *upper-body* can be decomposed into {*head*, *torso*, *upper-arm*, *lower-arm*}. Formally, for a node u, let us denote its child node set as $\mathcal{C}_u$. Decompositional relation has two properties: **(1)** the parent node and its child nodes are 'compatible', a pixel can be assigned by a label u and one from $\mathcal{C}_u$; and **(2)** the child nodes are mutually exclusive, a pixel can be assigned only one label from $\mathcal{C}_u$. Our decompositional relation function R^{dec} is designed as:

$$\begin{aligned} \boldsymbol{h}_{u,v} &= R^{\text{dec}}(F^{\text{dec}}_{u,v}(\boldsymbol{h}_u), \boldsymbol{h}_v), \quad v \in \mathcal{C}_u, \\ F^{\text{dec}}_{u,v}(\boldsymbol{h}_u) &= \boldsymbol{h}_u \odot \texttt{att}^{\text{dec}}_{u,v}(\boldsymbol{h}_u) \odot A(\boldsymbol{h}_u). \end{aligned} \tag{5}$$

Here $\texttt{att}^{\text{dec}}_{u,v}(\boldsymbol{h}_u) \in [0, 1]^{W\times H}$ is an attention map, which is computed for each sub-node $v \in \mathcal{C}_u$ and encodes the second constraint:

$$\begin{aligned} \texttt{att}^{\text{dec}}_{u,v}(\boldsymbol{h}_u) &= \text{PSM}([\phi_v(\boldsymbol{h}_u)]_{v\in\mathcal{C}_u}) \\ &= \frac{\exp(\phi_v(\boldsymbol{h}_u))}{\Sigma_{v'\in\mathcal{C}_u}\exp(\phi_{v'}(\boldsymbol{h}_u))}, \end{aligned} \tag{6}$$

where $\phi_v(\boldsymbol{h}_u) \in \mathbb{R}^{H\times W}$ computes a specific significance map for each child node v, PSM$(\cdot)$ represents *pixel-wise soft-max* and '$[\cdot]$' stands for concatenation. By making $\sum_{v\in\mathcal{C}_u}\texttt{att}^{\text{dec}}_{u,v} = \mathbf{1}$, $\{\texttt{att}^{\text{dec}}_{u,v}(\boldsymbol{h}_u)\}_{v\in\mathcal{C}_u}$ forms a *decompositional attention* mechanism, *i.e.*, allocates disparate attentions on different parts of $\boldsymbol{h}_u$, describing the second prop-

erty. In this way, the parent node u passes separate high-level information to different child nodes. The term $A(\boldsymbol{h}_u)$ is used to addresses the first property. It enforces the compatibility between u and $\mathcal{C}_u$, *i.e.*, lets the decompositional attentions only focus on the region within u, while suppresses the responses from other regions.

Concerning $A^{\text{dec}}_{u,v}$, please note that it is node-specific and separately learnt for the three entire nodes, namely *full-body*, *upper-body* and *lower-body*. Subscript u, v is added to address this point.

Compositional relation modeling: In the human hierarchy $\mathcal{G}$, compositional relations are represented by those vertical edges whose directions are opposite to the ones of the decompositional edges. To address this kind of relations, we design a compositional relation network R^{com} as:

$$\begin{aligned} \boldsymbol{h}_{u,v} &= R^{\text{com}}(F^{\text{com}}(\boldsymbol{h}_u), \boldsymbol{h}_v), \quad u \in \mathcal{C}_v, \\ F^{\text{com}}(\boldsymbol{h}_u) &= \boldsymbol{h}_u \odot \texttt{att}^{\text{com}}(\boldsymbol{h}_u) = \boldsymbol{h}_u \odot \text{PMP}([A(\boldsymbol{h}_{u'})]_{u' \in \mathcal{C}_v}), \end{aligned} \tag{7}$$

where $\text{PMP}(\cdot)$ stands for *pixel-wise max-pooling*. For node v, the statics from all the child nodes $\mathcal{C}_v$ are gathered for forming a non-parametric *compositional attention* $\texttt{att}^{\text{com}} \in [0,1]^{W \times H}$.

Iterative inference over human hierarchy: Human body presents a hierarchical structure. According to the graph theory, approximate inference algorithms must be used for such loopy structures $\mathcal{G}$. However, previous structured human parsers directly produce the final node representation $\boldsymbol{h}_v$ by accounting for contextual information from the parent node u: $\boldsymbol{h}_v \leftarrow R(\boldsymbol{h}_u, \boldsymbol{h}_v)$, where $v \in \mathcal{C}_u$; or using information from its neighbors $\mathcal{N}_v$: $\boldsymbol{h}_v \leftarrow \sum_{u \in \mathcal{N}_v} R(\boldsymbol{h}_u, \boldsymbol{h}_v)$. They neglect the fact that, in such structured human parsing setting, information is organized in a complex, cyclic system. The most generally used solutions are instead built upon iterative algorithms, *i.e.*, the node representation should be updated iteratively by aggregating the messages from its connected neighbors and the representation after several update iterations can approximate the optimal results [45]. In graph theory parlance, the iterative algorithm can be achieved by a parametric message passing process. It is defined in terms of a message function M and node update function U, and runs T steps. For each node v, the message passing process recursively collects information (messages) $\boldsymbol{m}_v$ from the neighbors $\mathcal{N}_v$ and updates the node state $\boldsymbol{h}_v$:

$$\begin{aligned} \boldsymbol{m}_v^{(t)} &= \sum\nolimits_{u \in \mathcal{N}_v} M(\boldsymbol{h}_u^{(t-1)}, \boldsymbol{h}_v^{(t-1)}), \\ \boldsymbol{h}_v^{(t)} &= U(\boldsymbol{h}_v^{(t-1)}, \boldsymbol{m}_v^{(t)}), \end{aligned} \tag{8}$$

where $\boldsymbol{h}_v^{(t)}$ stands for v's state in the t-th iteration and fully-connected recurrent neural networks are typically used to address the iterative nature of the update function U.

Inspired by previous message passing algorithms, our iterative algorithm is designed as below:

$$\begin{aligned} \boldsymbol{m}_v^{(t)} &= \underbrace{\sum\nolimits_{u \in \mathcal{K}_v} \boldsymbol{h}_{u,v}}_{\text{dependency}} + \underbrace{\sum\nolimits_{u \in \mathcal{P}_v} \boldsymbol{h}_{u,v}}_{\text{decomposition}} + \underbrace{\sum\nolimits_{u \in \mathcal{C}_v} \boldsymbol{h}_{u,v}}_{\text{composition}}, \\ \boldsymbol{h}_v^{(t)} &= U_{\text{convGRU}}(\boldsymbol{h}_v^{(t-1)}, \boldsymbol{m}_v^{(t)}), \end{aligned} \tag{9}$$

where the initial state $\boldsymbol{h}_v^{(0)}$ is obtained by Eq. 1. Here the message aggregation step is achieved by per-edge relation function terms, *i.e.*, the node v updates its state $\boldsymbol{h}_v$ by absorbing all the incoming information along different relations. As for the update function U, we use a convGRU, which replaces the fully-connected units in GRU with convolution operations, to describe its repeated activation behavior and address the pixel-wise nature of human parsing, simultaneously.

Loss function: In each step t, predictions $\hat{\mathcal{Y}}^{(t)} = \{\hat{y}_v^{(t)}\}_{v \in \mathcal{V}}$ of nodes $\mathcal{V}$ are obtained by applying a convolutional readout function $O : \mathbb{R}_{\geq 0}^{W \times H \times c} \mapsto \mathbb{R}^{W \times H}$ over $\{\boldsymbol{h}_v^{(t)}\}_{v \in \mathcal{V}}$:

$$\hat{\mathcal{Y}}^{(t)} = \{\hat{y}_v^{(t)}\}_{v \in \mathcal{V}} = \{O(\boldsymbol{h}_v^{(t)})\}_{v \in \mathcal{V}}. \tag{10}$$

Given the node outputs $\hat{\mathcal{Y}}^{(t)}$ and the corresponding human parsing groundtruths $\mathcal{Y}$, the learning task can be posed as the minimization of the following loss:

$$L_{\text{overall}} = \sum\nolimits_{t=1}^{T} L(\hat{\mathcal{Y}}^{(t)}, \mathcal{Y}), \tag{11}$$

where L stands for the standard cross entropy loss widely used for semantic segmentation.

3.3. Implementation Details

XXX

4. Experiments

Herein, we describe our experimental settings (§4.1), report quantitative results comparing to several state-of-the-arts on five datasets (∼20K test images in total, §4.2), show qualitative results (§4.3), and study the impact of different components of our model (§4.4). All the visual results shown in this section are drawn from the test sets. More quantitative and qualitative results are provided in the supplementary material.

4.1. Experimental Setting

Datasets: Five standard benchmark datasets [20, 50, 39, 28, 40] are used for a comprehensive performance evaluation. LIP [20] contains 50,462 single-person images, which are collected from realistic scenarios and divided into 30,462 for training, 10,000 for validation and 10,000 for test. The pixel-wise annotations cover 19 human part categories (*e.g.*, *face*, *left-/right-arms*, *left-/right-legs*, *etc.*). PASCAL-Person-Part [50] includes 3,533 multi-person images with challenging poses and viewpoints. Each image is pixel-wise annotated with six classes (*i.e.*, *head, torso, upper-/lower-arms*, and *upper-/lower-legs*). It is split into 1,716 and 1,817 images for training and test. ATR [28] is a

Methods	pixAcc.	Mean Acc.	Mean IoU
SegNet [1]	69.04	24.00	18.17
FCN-8s [37]	76.06	36.75	28.29
DeepLabV2 [5]	82.66	51.64	41.64
Attention [6]	83.43	54.39	42.92
†Attention+SSL [20]	84.36	54.94	44.73
DeepLabV3+ [7]	84.09	55.62	44.80
ASN [38]	-	-	45.41
†SSL [20]	-	-	46.19
MMAN [41]	-	-	46.81
†SS-NAN [58]	87.59	56.03	47.92
HSP-PRI [22]	85.07	60.54	48.16
†MuLA [42]	**88.5**	60.5	49.3
PSPNet [56]	86.23	61.33	50.56
CE2P [35]	87.37	63.20	53.10
BraidNet [36]	87.60	66.09	54.42
CNIF [48]	88.03	68.80	57.74
Ours	XX	**XX**	**58.96**

Table 1: **Comparison of pixel accuracy, mean accuracy and mIoU on LIP `val` [20].** † indicates extra pose information used. (Higher values are better. The best score is marked in **bold**. These notes are the same for other tables.)

challenging human parsing dataset, which has 7,700 single-person images with dense annotations over 17 categories (*e.g.*, *face, upper-clothes, left-/right-arms, left-/right-legs*, *etc.*). There are 6,000/700/1,000 images for training, validation, and test, respectively. PPSS [39] collects 3,673 single-pedestrian images from 171 surveillance videos and provides pixel-wise annotations for *hair, face, upper-/lower-clothes, arm*, and *leg*. It presents diverse real-word challenges, *e.g.*, pose variants, illumination changes, and occlusions. There are 1,781 and 1,892 images for training and testing, respectively. Fashion Clothing [40] has 4,371 images gathered from Colorful Fashion Parsing [32], Fashionista [53], and Clothing Co-Parsing [54]. It has 17 clothing categories (*e.g.*, *hair, pants, shoes, upper-clothes*, *etc.*) and the data split follows 3,934 for training and 437 for test.

Reproducibility: Our method is implemented on PyTorch and trained on four NVIDIA Tesla V100 GPUs with 32GB memory per-card. All the experiments are performed on one NVIDIA TITAN Xp 12GB GPU. To provide full details of our approach, our code will be made publicly available.

Evaluation: For fair comparison, we follow the official evaluation protocols of each dataset. For LIP, following [58], we report pixel accuracy, mean accuracy and mean Intersection-over-Union (mIoU). For PASCAL-Person-Part and PPSS, following [49, 50, 41], the performance is evaluated in terms of mIoU. For ATR and Fashion Clothing, as in [40, 48], we report pixel accuracy, foreground accuracy, average precision, average recall, and average F1-score.

4.2. Quantitative Results

LIP [20]: LIP is one of gold standard benchmarks for human parsing. We report the results in Table 1 to compare our method with 13 recent state-of-the-arts on LIP `val` set. We also would like to mention that our networks do not use extra pose information such as pose or edge.

PASCAL-Person-Part [50]: In Table 3, we compare our method with 15 state-of-the-arts on its `test` set using IoU score. From the results, we can see that our approach achieves better performance compared to all other methods. For example, our approach achieves XX% mIoU, which improves the recent XX by 0.2%. Especially, improving 0.2% is not neglectable considering the improvement on PASCAL-Person-Part is very challenging.

ATR [28]: We compare our approach to the previous state-of-the-arts on ATR `test` set and illustrate the results in Table 4. According to the results, our approach achieves new state-of-the-art performance XX%, which outperforms all the other methods by a large margin.

Fashion Clothing [40]: As reported in Table 5, we achieve 59.77% F-score on Fashion Clothing `test`.

PPSS [39]: As illustrated in Table 6, we compare our approach with XXX previous state-of-the-arts on PPSS `test` set.

4.3. Qualitative Results

XXX

4.4. Ablation Study

5. Conclusion

XXX

References

[1] V. Badrinarayanan, A. Kendall, and R. Cipolla. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE TPAMI*, 39(12):2481–2495, 2017. 5, 6

[2] Y. Bo and C. C. Fowlkes. Shape-based pedestrian parsing. In *CVPR*, pages 2265–2272, 2011. 2

[3] H. Chen, Z. J. Xu, Z. Q. Liu, and S. C. Zhu. Composite templates for cloth modeling and sketching. In *CVPR*, pages 943–950, 2006. 2

[4] L.-C. Chen, M. Collins, Y. Zhu, G. Papandreou, B. Zoph, F. Schroff, H. Adam, and J. Shlens. Searching for efficient multi-scale architectures for dense image prediction. In *NIPS*, pages 8699–8710, 2018. 6

[5] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE TPAMI*, 40(4):834–848, 2018. 1, 5, 6

[6] 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. 1, 5, 6

[7] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *ECCV*, 2018. 5, 6

[8] 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

[9] B. Cheng, L.-C. Chen, Y. Wei, Y. Zhu, Z. Huang, J. Xiong, T. Huang, W.-M. Hwu, and H. Shi. Spgnet: Semantic prediction guidance for scene parsing. In *ICCV*, 2019. 6

Methods	Hat	Hair	Glov	Sung	Clot	Dress	Coat	Sock	Pant	Suit	Scarf	Skirt	Face	L-Arm	R-Arm	L-Leg	R-Leg	L-Sh	R-Sh	B.G.	Ave.
SegNet [1]	26.60	44.01	0.01	0.00	34.46	0.00	15.97	3.59	33.56	0.01	0.00	0.00	52.38	15.30	24.23	13.82	13.17	9.26	6.47	70.62	18.17
FCN-8s [37]	39.79	58.96	5.32	3.08	49.08	12.36	26.82	15.66	49.41	6.48	0.00	2.16	62.65	29.78	36.63	28.12	26.05	17.76	17.70	78.02	28.29
DeepLabV2 [5]	56.48	65.33	29.98	19.67	62.44	30.33	51.03	40.51	69.00	22.38	11.29	20.56	70.11	49.25	52.88	42.37	35.78	33.81	32.89	84.53	41.64
Attention [6]	58.87	66.78	23.32	19.48	63.20	29.63	49.70	35.23	66.04	24.73	12.84	20.41	70.58	50.17	54.03	38.35	37.70	26.20	27.09	84.00	42.92
Attention+SSL [20]	59.75	67.25	28.95	21.57	65.30	29.49	51.92	38.52	68.02	24.48	14.92	24.32	71.01	52.64	55.79	40.23	38.80	28.08	29.03	84.56	44.73
ASN [38]	56.92	64.34	28.07	17.78	64.90	30.85	51.90	39.75	71.78	25.57	7.97	17.63	70.77	53.53	56.70	49.58	48.21	34.57	33.31	84.01	45.41
SSL [20]	58.21	67.17	31.20	23.65	63.66	28.31	52.35	39.58	69.40	28.61	13.70	22.52	74.84	52.83	55.67	48.22	47.49	31.80	29.97	84.64	46.19
MMAN [41]	57.66	65.63	30.07	20.02	64.15	28.39	51.98	41.46	71.03	23.61	9.65	23.20	69.54	55.30	58.13	51.90	52.17	38.58	39.05	84.75	46.81
SS-NAN [58]	63.86	70.12	30.63	23.92	70.27	33.51	56.75	40.18	72.19	27.68	16.98	26.41	75.33	55.24	58.93	44.01	41.87	29.15	32.64	88.67	47.92
CE2P [35]	65.29	72.54	39.09	32.73	69.46	32.52	56.28	49.67	74.11	27.23	14.19	22.51	75.50	65.14	66.59	60.10	58.59	46.63	46.12	87.67	53.10
BraidNet	66.8	72.0	42.5	32.1	69.8	33.7	57.4	49.0	74.9	32.4	19.3	27.2	74.9	65.5	67.9	60.2	59.6	47.4	47.9	88.0	54.4
CNIF [48]	69.55	73.45	45.17	41.45	70.57	38.52	57.94	54.02	75.07	28.00	31.92	30.20	76.38	68.28	69.49	65.52	65.51	52.67	53.38	87.99	57.74
Ours	**70.56**	**75.01**	**46.32**	**42.68**	**71.77**	**40.20**	**59.03**	**55.21**	**76.35**	**29.51**	**33.40**	**31.69**	**77.13**	**69.34**	**70.38**	**66.50**	**66.47**	**53.84**	**54.52**	**89.27**	**58.96**

Table 2: **Per-class comparison of mIoU with state-of-the-art methods on LIP `val` [20].**

Methods	Head	Torso	U-Arm	L-Arm	U-Leg	L-Leg	B.G.	Ave.
HAZN [49]	80.79	59.11	43.05	42.76	38.99	34.46	93.59	56.11
Attention [6]	81.47	59.06	44.15	42.50	38.28	35.62	93.65	56.39
LG-LSTM [30]	82.72	60.99	45.40	47.76	42.33	37.96	88.63	57.97
Attention+SSL [20]	83.26	62.40	47.80	45.58	42.32	39.48	94.68	59.36
Attention+MMAN [41]	82.58	62.83	48.49	47.37	42.80	40.40	94.92	59.91
Graph LSTM [29]	82.69	62.68	46.88	47.71	45.66	40.93	94.59	60.16
SS-NAN [58]	86.43	67.28	51.09	48.07	44.82	42.15	97.23	62.44
Structure LSTM [27]	82.89	67.15	51.42	48.72	51.72	45.91	97.18	63.57
Joint [50]	85.50	67.87	54.72	54.30	48.25	44.76	95.32	64.39
DeepLabV2 [5]	-	-	-	-	-	-	-	64.94
MuLA [42]	84.6	68.3	57.5	54.1	49.6	46.4	95.6	65.1
PCNet [59]	86.81	69.06	55.35	55.27	50.21	48.54	96.07	65.90
Holistic [26]	86.00	69.85	56.63	55.92	51.46	48.82	95.73	66.34
WSHP [16]	87.15	72.28	57.07	56.21	52.43	50.36	**97.72**	67.60
DeepLabV3+ [7]	87.02	72.02	60.37	57.36	53.54	48.52	96.07	67.84
SPGNet [9]	87.67	71.41	61.69	60.35	52.62	48.80	95.98	68.36
PGN [19]	**90.89**	**75.12**	55.83	64.61	55.42	41.57	95.33	68.40
CNIF [48]	88.02	72.91	64.31	63.52	55.61	54.96	96.02	70.76
Graphonomy [18]	88.04	74.99	61.49	60.72	54.21	50.06	95.87	71.14
DPC [4]	88.81	74.54	63.85	63.73	57.24	54.55	96.66	71.34
Ours	89.12	74.35	**65.86**	**65.03**	**57.46**	**56.87**	96.51	**72.17**

Table 3: **Per-class comparison of mIoU with state-of-the-art methods on PASCAL-Person-Part `test` [50].**

Methods	pixAcc.	F.G. Acc.	Prec.	Recall	F-1
Yamaguchi [53]	84.38	55.59	37.54	51.05	41.80
Paperdoll [52]	88.96	62.18	52.75	49.43	44.76
M-CNN [33]	89.57	73.98	64.56	65.17	62.81
ATR [28]	91.11	71.04	71.69	60.25	64.38
DeepLabV2 [5]	94.42	82.93	78.48	69.24	73.53
PSPNet [56]	95.20	80.23	79.66	73.79	75.84
Attention [6]	95.41	85.71	81.30	73.55	77.23
DeepLabV3+ [7]	95.96	83.04	80.41	78.79	79.49
Co-CNN [31]	96.02	83.57	84.95	77.66	80.14
LG-LSTM [30]	96.18	84.79	84.64	79.43	80.97
TGPNet [40]	96.45	87.91	83.36	80.22	81.76
Graph LSTM [29]	97.60	91.42	84.74	83.28	83.76
CNIF [48]	96.26	87.91	84.62	86.41	85.51
Structure LSTM [27]	97.71	91.76	89.37	86.84	87.88
Graphonomy [18]	98.32	-	-	-	90.89
Ours	**96.53**	**88.75**	**85.72**	**87.96**	**86.82**

Table 4: **Comparison of accuracy, foreground accuracy, average precision, recall and F1-score on ATR `test` [28].** Please see the supplementary material for per-class performance.

Methods	pixAcc.	F.G. Acc.	Prec.	Recall	F-1
Yamaguchi [53]	81.32	32.24	23.74	23.68	22.67
Paperdoll [52]	87.17	50.59	45.80	34.20	35.13
DeepLabV2 [5]	87.68	56.08	35.35	39.00	37.09
Attention [6]	90.58	64.47	47.11	50.35	48.68
TGPNet [40]	91.25	66.37	50.71	53.18	51.92
CNIF [48]	92.20	68.59	56.84	59.47	58.12
Ours	**92.75**	**69.88**	**58.29**	**61.32**	**59.77**

Table 5: **Comparison of pixel accuracy, foreground pixel accuracy, average precision, average recall and average f1-score on Fashion Clothing `test` [40].**

Methods	Head	Face	U-Cloth	Arms	L-Cloth	Legs	B.G.	Ave.
DL [39]	22.0	29.1	57.3	10.6	46.1	12.9	68.6	35.2
DDN [39]	35.5	44.1	68.4	17.0	61.7	23.8	80.0	47.2
ASN [38]	51.7	51.0	65.9	29.5	52.8	20.3	83.8	50.7
MMAN [41]	53.1	50.2	69.0	29.4	55.9	21.4	85.7	52.1
LCPC [11]	55.6	46.6	71.9	30.9	58.8	24.6	86.2	53.5
CNIF [48]	67.6	60.8	80.8	46.8	69.5	28.7	90.6	60.5
Ours	**68.5**	**62.7**	**81.4**	**48.8**	**70.6**	**31.2**	**91.1**	**64.9**

Table 6: **Comparison of mIoU on PPSS `test` [39].**

for image parsing with regular spatial layout. In *BMVC*, 2014. 6

[12] J. Dong, Q. Chen, X. Shen, J. Yang, and S. Yan. Towards unified human parsing and pose estimation. In *CVPR*, pages 843–850, 2014. 2

[13] J. Dong, Q. Chen, W. Xia, Z. Huang, and S. Yan. A deformable mixture parsing model with parselets. In *ICCV*, pages 3408–3415, 2013. 2

[14] D. K. Duvenaud, D. Maclaurin, J. Iparraguirre, R. Bombarell, T. Hirzel, A. Aspuru-Guzik, and R. P. Adams. Convolutional networks on graphs for learning molecular fingerprints. In *NIPS*, 2015. 2

[15] S. Eslami and C. Williams. A generative model for parts-based object segmentation. In *NIPS*, pages 100–107, 2012. 2

[16] H.-S. Fang, G. Lu, X. Fang, J. Xie, Y.-W. Tai, and C. Lu. Weakly and semi supervised human body part parsing via pose-guided knowledge transfer. In *CVPR*, pages 70–78, 2018. 1, 2, 6

[17] J. Gilmer, S. S. Schoenholz, P. F. Riley, O. Vinyals, and G. E. Dahl. Neural message passing for quantum chemistry. In *ICML*, pages 1263–1272, 2017. 1, 2, 3

[18] K. Gong, Y. Gao, X. Liang, X. Shen, M. Wang, and L. Lin.

Graphonomy: Universal human parsing via graph transfer learning. In *CVPR*, 2019. 2, 6

[19] K. Gong, X. Liang, Y. Li, Y. Chen, M. Yang, and L. Lin. Instance-level human parsing via part grouping network. In *ECCV*, pages 770–785, 2018. 1, 6

[20] K. Gong, X. Liang, D. Zhang, X. Shen, and L. Lin. Look into person: Self-supervised structure-sensitive learning and a new benchmark for human parsing. In *CVPR*, pages 932–940, 2017. 1, 2, 4, 5, 6

[21] W. L. Hamilton, R. Ying, and J. Leskovec. Representation learning on graphs: Methods and applications. *arXiv preprint arXiv:1709.05584*, 2017. 2

[22] M. M. Kalayeh, E. Basaran, M. Gokmen, M. E. Kamasak, and M. Shah. Human semantic parsing for person re-identification. In *CVPR*, 2018. 5

[23] S. Kearnes, K. McCloskey, M. Berndl, V. Pande, and P. Riley. Molecular graph convolutions: moving beyond fingerprints. *Journal of computer-aided molecular design*, 30(8):595–608, 2016. 3

[24] T. N. Kipf and M. Welling. Semi-supervised classification with graph convolutional networks. In *ICLR*, 2017. 2

[25] L. Ladicky, P. H. Torr, and A. Zisserman. Human pose estimation using a joint pixel-wise and part-wise formulation. In *CVPR*, pages 3578–3585, 2013. 2

[26] Q. Li, A. Arnab, and P. H. Torr. Holistic, instance-level human parsing. In *BMVC*, 2017. 6

[27] X. Liang, L. Lin, X. Shen, J. Feng, S. Yan, and E. P. Xing. Interpretable structure-evolving lstm. In *CVPR*, pages 1010–1019, 2017. 6

[28] X. Liang, S. Liu, X. Shen, J. Yang, L. Liu, J. Dong, L. Lin, and S. Yan. Deep human parsing with active template regression. *IEEE TPAMI*, 37(12):2402–2414, 2015. 1, 2, 4, 5, 6

[29] X. Liang, X. Shen, J. Feng, L. Lin, and S. Yan. Semantic object parsing with graph lstm. In *ECCV*, pages 125–143, 2016. 2, 6

[30] X. Liang, X. Shen, D. Xiang, J. Feng, L. Lin, and S. Yan. Semantic object parsing with local-global long short-term memory. In *CVPR*, pages 3185–3193, 2016. 2, 6

[31] 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, 6

[32] S. Liu, J. Feng, C. Domokos, H. Xu, J. Huang, Z. Hu, and S. Yan. Fashion parsing with weak color-category labels. *TMM*, 16(1):253–265, 2014. 2, 5

[33] S. Liu, X. Liang, L. Liu, X. Shen, J. Yang, C. Xu, L. Lin, X. Cao, and S. Yan. Matching-cnn meets knn: Quasi-parametric human parsing. In *CVPR*, pages 1419–1427, 2015. 2, 6

[34] S. Liu, Y. Sun, D. Zhu, G. Ren, Y. Chen, J. Feng, and J. Han. Cross-domain human parsing via adversarial feature and label adaptation. In *AAAI*, 2018. 2

[35] T. Liu, T. Ruan, Z. Huang, Y. Wei, S. Wei, Y. Zhao, and T. Huang. Devil in the details: Towards accurate single and multiple human parsing. *arXiv preprint arXiv:1809.05996*, 2018. 5, 6

[36] X. Liu, M. Zhang, W. Liu, J. Song, and T. Mei. Braidnet: Braiding semantics and details for accurate human parsing. In *ACMMM*, 2019. 5

[37] J. Long, E. Shelhamer, and T. Darrell. Fully convolutional networks for semantic segmentation. In *CVPR*, pages 3431–3440, 2015. 1, 5, 6

[38] P. Luc, C. Couprie, S. Chintala, and J. Verbeek. Semantic segmentation using adversarial networks. In *NIPS-workshop*, 2016. 5, 6

[39] P. Luo, X. Wang, and X. Tang. Pedestrian parsing via deep decompositional network. In *ICCV*, pages 2648–2655, 2013. 1, 2, 4, 5, 6

[40] X. Luo, Z. Su, J. Guo, G. Zhang, and X. He. Trusted guidance pyramid network for human parsing. In *ACMMM*, 2018. 1, 2, 4, 5, 6

[41] Y. Luo, Z. Zheng, L. Zheng, T. Guan, J. Yu, and Y. Yang. Macro-micro adversarial network for human parsing. In *ECCV*, pages 418–434, 2018. 2, 5, 6

[42] X. Nie, J. Feng, and S. Yan. Mutual learning to adapt for joint human parsing and pose estimation. In *ECCV*, pages 502–517, 2018. 1, 2, 5, 6

[43] M. Niepert, M. Ahmed, and K. Kutzkov. Learning convolutional neural networks for graphs. In *ICML*, pages 2014–2023, 2016. 2

[44] I. Rauschert and R. T. Collins. A generative model for simultaneous estimation of human body shape and pixel-level segmentation. In *ECCV*, pages 704–717, 2012. 2

[45] F. Scarselli, M. Gori, A. C. Tsoi, M. Hagenbuchner, and G. Monfardini. The graph neural network model. *IEEE Transactions on Neural Networks*, 20(1):61–80, 2008. 2, 3, 4

[46] P. Velickovic, G. Cucurull, A. Casanova, A. Romero, P. Lio, and Y. Bengio. Graph attention networks. In *ICLR*, 2018. 1, 2

[47] N. Wang and H. Ai. Who blocks who: Simultaneous clothing segmentation for grouping images. In *ICCV*, pages 1535–1542, 2011. 2

[48] W. Wang, Z. Zhang, S. Qi, J. Shen, Y. Pang, and L. Shao. Learning compositional neural information fusion for human parsing. In *ICCV*, 2019. 1, 2, 3, 5, 6

[49] 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. 2, 5, 6

[50] F. Xia, P. Wang, X. Chen, and A. L. Yuille. Joint multi-person pose estimation and semantic part segmentation. In *CVPR*, pages 6769–6778, 2017. 1, 2, 4, 5, 6

[51] F. Xia, J. Zhu, P. Wang, and A. L. Yuille. Pose-guided human parsing by an and/or graph using pose-context features. In *AAAI*, 2016. 1, 2

[52] K. Yamaguchi, M. Hadi Kiapour, and T. L. Berg. Paper doll parsing: Retrieving similar styles to parse clothing items. In *ICCV*, pages 3519–3526, 2013. 2, 6

[53] K. Yamaguchi, M. H. Kiapour, L. E. Ortiz, and T. L. Berg. Parsing clothing in fashion photographs. In *CVPR*, pages 3570–3577, 2012. 2, 5, 6

[54] W. Yang, P. Luo, and L. Lin. Clothing co-parsing by joint image segmentation and labeling. In *CVPR*, pages 3182–3189, 2014. 2, 5

[55] S. Zagoruyko and N. Komodakis. Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer. In *ICLR*, 2017. 3

[56] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia. Pyramid scene parsing network. In *CVPR*, pages 6230–6239, 2017. 5, 6

[57] J. Zhao, J. Li, Y. Cheng, T. Sim, S. Yan, and J. Feng. Under-

standing humans in crowded scenes: Deep nested adversarial learning and a new benchmark for multi-human parsing. In *ACMMM*, pages 792–800, 2018. 2

[58] 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 *CVPR-workshop*, pages 7–15, 2017. 1, 2, 5, 6

[59] B. Zhu, Y. Chen, M. Tang, and J. Wang. Progressive cognitive human parsing. In *AAAI*, 2018. 1, 2, 3, 6