Hierarchical Human Parsing with Typed Part-Relation Reasoning

Wenguan Wang[1,2*], Hailong Zhu[3*], Jifeng Dai[4], Yanwei Pang[3†], Jianbing Shen[2], Ling Shao[2]

[1] ETH Zurich, Switzerland [2] Inception Institute of Artificial Intelligence, UAE

[3] Tianjin Key Laboratory of Brain-inspired Intelligence Technology, School of Electrical and Information Engineering, Tianjin University, China

[4] SenseTime Research

wenguanwang.ai@gmail.com, hlzhu2009@gmail.com

https://github.com/hlzhu09/Hierarchical-Human-Parsing

Abstract

Human parsing is for pixel-wise human semantic understanding. As human bodies are underlying hierarchically structured, how to model human structures is the central theme in this task. Focusing on this, we seek to simultaneously exploit the representational capacity of deep graph networks and the hierarchical human structures. In particular, we provide following two contributions. First, three kinds of part relations, i.e.*, decomposition, composition, and dependency, are, for the first time, completely and precisely described by three distinct relation networks. This is in stark contrast to previous parsers, which only focus on a portion of the relations and adopt a type-agnostic relation modeling strategy. More expressive relation information can be captured by explicitly imposing the parameters in the relation networks to satisfy the specific characteristics of different relations. Second, previous parsers largely ignore the need for an approximation algorithm over the loopy human hierarchy, while we instead address an iterative reasoning process, by assimilating generic message-passing networks with their edge-typed, convolutional counterparts. With these efforts, our parser lays the foundation for more sophisticated and flexible human relation patterns of reasoning. Comprehensive experiments on five datasets demonstrate that our parser sets a new state-of-the-art on each.*

1. Introduction

Human parsing involves segmenting 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 finds a wide spectrum of human-centric applications, such as human behavior analysis [50, 58, 14], human-robot interaction [16], *etc*.

Human bodies present a highly structured hierarchy and body parts inherently interact with each other. As shown in Fig. 1(b), there are different relations between parts [42, 60, 49]: **decompositional** and **compositional** relations (full line:/) between constituent and entire parts (*e.g.*, {*upper body*, *lower body*} and *full body*), and **dependency** relations (dashed line:--) between kinematically connected parts (*e.g.*, *hand* and *arm*). Thus the central problem in human parsing is how to model such relations. Recently, numerous structured human parsers have been proposed [65, 15, 22, 64, 47, 74, 61, 20]. Their notable successes indeed demonstrate the benefit of exploiting the structure in this problem. However, three major limitations in human structure modeling are still observed. **(1)** The structural information utilized is typically weak and relation types studied are incomplete. Most efforts [65, 15, 22, 64, 47] directly encode human pose information into the parsing model, causing them to suffer from trivial structural information, not to mention the need of extra pose annotations. In addition, previous structured parsers focus on only one or two of the aforementioned part relations, not all of them. For example, [20] only considers

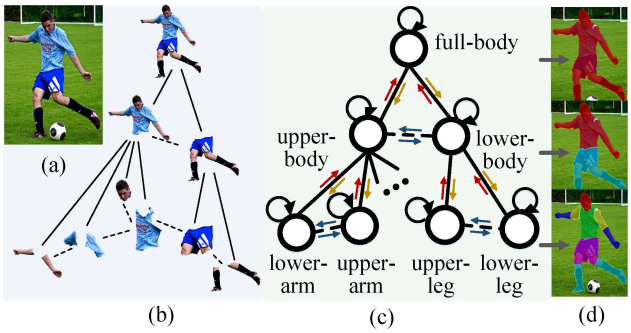


Figure 1: **Illustration of our hierarchical human parser.** (a) Input image. (b) The human hierarchy in (a), where -- indicates dependency relations and / is de-/compositional relations. (c) In our parser, three distinct relation networks are designed for addressing the specific characteristics of different part relations, *i.e.*, →, →, and → stand for decompositional, compositional, and dependency relation networks, respectively. Iterative inference (↻) is performed for better approximation. For visual clarity, some nodes are omitted. (d) Our hierarchical parsing results.

*The first two authors contribute equally to this work.
†Corresponding author: *Yanwei Pang*.

dependency relations, and [74] relies on decompositional relations. **(2)** Only a single relation model is learnt to reason different kinds of relations, without considering their essential and distinct geometric constraints. Such a relation modeling strategy is over-general and simple; do not seem to characterize well the diverse part relations. **(3)** According to graph theory, as the human body yields a complex, cyclic topology, an iterative inference is desirable for optimal result approximation. However, current arts [22, 64, 47, 74, 61] are primarily built upon an immediate, feed-forward prediction scheme.

To respond to the above challenges and enable a deeper understanding of human structures, we develop a unified, structured human parser that precisely describes a more complete set of part relations, and efficiently reasons structures with the prism of a message-passing, feed-back inference scheme. To address the first two issues, we start with an in-depth and comprehensive analysis on three essential relations, namely decomposition, composition, and dependency. Three distinct relation networks (→, →, and → in Fig. 1(c)) are elaborately designed and imposed to explicitly satisfy the specific, intrinsic relation constraints. Then, we construct our parser as a tree-like, end-to-end trainable graph model, where the nodes represent the human parts, and edges are built upon the relation networks. For the third issue, a modified, relation-typed convolutional message passing procedure (↻ in Fig. 1(c)) is performed over the human hierarchy, enabling our method to obtain better parsing results from a global view. All components, *i.e.*, the part nodes, edge (relation) functions, and message passing modules, are fully differentiable, enabling our whole framework to be end-to-end trainable and, in turn, facilitating learning about parts, relations, and inference algorithms.

More crucially, our structured human parser can be viewed as an essential variant of message passing neural networks (MPNNs) [19, 56], yet significantly differentiated in two aspects. (1) Most previous MPNNs are edge-type-agnostic, while ours addresses relation-typed structure reasoning with a higher expressive capability. (2) By replacing the Multilayer Perceptron (MLP) based MPNN units with convolutional counterparts, our parser gains a spatial information preserving property, which is desirable for such a pixel-wise prediction task.

We extensively evaluate our approach on five standard human parsing datasets [22, 64, 44, 31, 45], achieving state-of-the-art performance on all of them (§4.2). In addition, with ablation studies for each essential component in our parser (§4.3), three key insights are found: (1) Exploring different relations reside on human bodies is valuable for human parsing. (2) Distinctly and explicitly modeling different types of relations can better support human structure reasoning. (3) Message passing based feed-back inference is able to reinforce parsing results.

2. Related Work

Human parsing: Over the past decade, active research has been devoted towards pixel-level human semantic understanding. Early approaches tended to leverage image regions [35, 68, 69], hand-crafted features [57, 7], part templates [2, 11, 10] and human keypoints [67, 35, 68, 69], and typically explored certain heuristics over human body configurations [3, 11, 10] in a CRF [67, 28], structured model [68, 11], grammar model [3, 42, 10], or generative model [13, 51] framework. Recent advance has been driven by the streamlined designs of deep learning architectures. Some pioneering efforts revisit classic template matching strategy [31, 36], address local and global cues [34], or use tree-LSTMs to gather structure information [32, 33]. However, due to the use of superpixel [34, 32, 33] or HOG feature [44], they are fragmentary and time-consuming. Consequent attempts thus follow a more elegant FCN architecture, addressing multi-level cues [5, 63], feature aggregation [45, 72, 38], adversarial learning [71, 46, 37], or cross-domain knowledge [37, 66, 20]. To further explore inherent structures, numerous approaches [65, 72, 22, 64, 15, 47] choose to straightforward encode pose information into the parsers, however, relying on off-the-shelf pose estimators [18, 17] or additional annotations. Some others consider top-down [74] or multi-source semantic [61] information over hierarchical human layouts. Though impressive, they ignore iterative inference 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 the different relations residing on human bodies, *i.e.*, decomposition, composition, and dependency, and addressing iterative, spatial-information preserving inference over human hierarchy.

Graph neural networks (GNNs):GNNs have a rich history (dating back to [53]) and became a veritable explosion in research community over the last few years [23]. GNNs effectively learn graph representations in an end-to-end manner, and can generally be divided into two broad classes: Graph Convolutional Networks (GCNs) and Message Passing Graph Networks (MPGNs). The former [12, 48, 27] directly extend classical CNNs to non-Euclidean data. Their simple architecture promotes their popularity, while limits their modeling capability for complex structures [23]. MPGNs [19, 73, 56, 59] parameterize all the nodes, edges, and 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 to explore GNNs in the area of human parsing. In contrast to conventional MPGNs, which are mainly MLP-based and edge-type-agnostic, we provide a spatial information preserving and relation-type aware graph learning scheme.

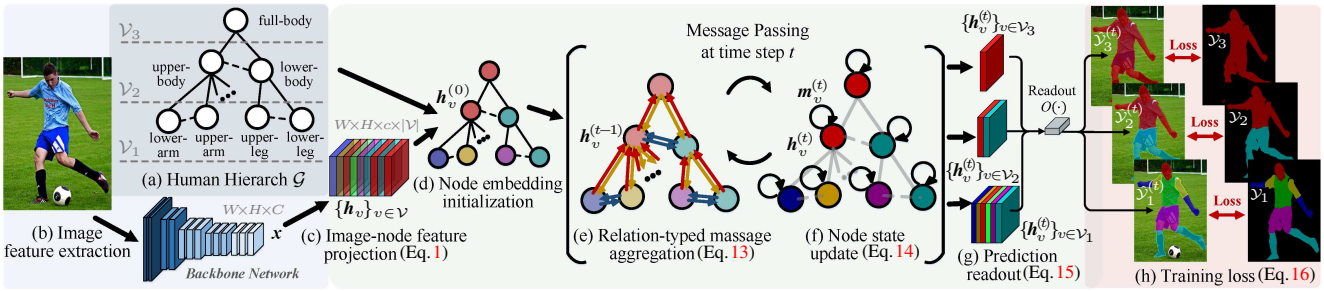


Figure 2: Illustration of our structured human parser for hierarchical human parsing during the training phase. The main components in the flowchart are marked by (a)-(h). Please refer to §3 for more details. Best viewed in color.

3. Our Approach

3.1. Problem Definition

Formally, we represent the human semantic structure as a directed, hierarchical graph $\mathcal{G} = (\mathcal{V}, \mathcal{E}, \mathcal{Y})$. As show in Fig. 2(a), the node set $\mathcal{V} = \cup_{l=1}^{3}\mathcal{V}_l$ represents human parts in three different semantic levels, including the leaf nodes $\mathcal{V}_1$ (*i.e.*, the most fine-grained parts: *head, arm, hand, etc.*) which are typically considered in common human parsers, two middle-level nodes $\mathcal{V}_2$=$\{$*upper-body, lower-body*$\}$ and one root $\mathcal{V}_3$ =$\{$*full-body*$\}$[1]. The edge set $\mathcal{E}\in\binom{\mathcal{V}}{2}$ represents the relations between 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 and each edge (u, v) are associated with feature vectors: $\boldsymbol{h}_v$ and $\boldsymbol{h}_{u,v}$, respectively. $y_v\in\mathcal{Y}$ indicates the groundtruth segmentation map of part (node) v and the groundtruth maps $\mathcal{Y}$ are also organized in a hierarchical manner: $\mathcal{Y}=\cup_{l=1}^{3}\mathcal{Y}_l$.

Our human parser is trained in a graph learning scheme, using the full supervision from existing human parsing datasets. For a test sample, it is able to effectively infer the node and edge representations by reasoning human structures at the levels of individual parts and their relations, and iteratively fusing the information over the human structures.

3.2. Structured Human Parsing Network

Node embedding: As an initial step, a learnable projection function is used to map the input image representation into node (part) features, in order to obtain sufficient expressive power. Formally, let us denote the input image feature as $\boldsymbol{x}\in\mathbb{R}^{W\times H\times C}$, which comes from a DeepLabV3 [6]-like backbone network (in Fig. 2(b)), and the projection function as P: $\mathbb{R}^{W\times H\times C}\mapsto\mathbb{R}^{W\times H\times c\times|\mathcal{V}|}$, where $|\mathcal{V}|$ indicates the number of nodes. The node embeddings $\{\boldsymbol{h}_v\in\mathbb{R}^{W\times H\times c}\}_{v\in\mathcal{V}}$ are initialized by (Fig. 2(d)):

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

where each node embedding $\boldsymbol{h}_v$ is a (W, H, c)-dimensional tenor that encodes full spatial details (in Fig. 2(c)).

Typed human part relation modeling: Basically, an edge embedding $\boldsymbol{h}_{u,v}$ captures the relations between nodes u and v. Most previous structured human parsers [74, 61] work in an edge-type-agnostic manner, *i.e.*, a unified, shared relation network R: $\mathbb{R}^{W\times H\times c}\times\mathbb{R}^{W\times H\times c}\mapsto\mathbb{R}^{W\times H\times c}$ is used to capture all the relations: $\boldsymbol{h}_{u,v}=R(\boldsymbol{h}_u,\boldsymbol{h}_v)$. Such a strategy may lose the discriminability of individual relation types and does not have an explicit bias towards modeling geometric and anatomical constraints. In contrast, we formulate $\boldsymbol{h}_{u,v}$ in a relation-typed manner R^r:

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

where $r \in \{\text{dec}, \text{com}, \text{dep}\}$. $F^r(\cdot)$ is an attention-based relation-adaption operation, which is used to enhance the original node embedding $\boldsymbol{h}_u$ by addressing geometric characteristics in relation r. The attention mechanism is favored here as it allows trainable and flexible feature enhancement and explicitly encodes specific relation constraints. From the view of information diffusion mechanism in the graph theory [53], if there exists an edge (u, v) that links a starting node u to a destination v, this indicates v should receive 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, employing the more tractable feature $F^r(\boldsymbol{h}_u)$ in place of $\boldsymbol{h}_u$, so more expressive relation feature $\boldsymbol{h}_{u,v}$ for v can be obtained and further benefit the final parsing results. In this way, we learn more sophisticated and impressive relation patterns within human bodies.

1) *Decompositional relation modeling:* Decompositional relations (full line:/ in Fig. 2(a)) are represented by those vertical edges starting from parent nodes to corresponding child nodes in the human hierarchy $\mathcal{G}$. For example, 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 $\mathcal{C}_u$. Our decompositional relation network R^{dec} aims to learn the rule for 'breaking down' u into its constituent parts $\mathcal{C}_u$ (Fig. 3):

$$\begin{aligned} \boldsymbol{h}_{u,v} &= R^{\text{dec}}(F^{\text{dec}}(\boldsymbol{h}_u), \boldsymbol{h}_v), \quad v \in \mathcal{C}_u, \\ F^{\text{dec}}(\boldsymbol{h}_u) &= \boldsymbol{h}_u \odot \texttt{att}^{\text{dec}}_{u,v}(\boldsymbol{h}_u). \end{aligned} \tag{3}$$

'$\odot$' indicates the attention-based feature enhancement operation, and $\texttt{att}^{\text{dec}}_{u,v}(\boldsymbol{h}_u) \in [0, 1]^{W\times H}$ produces an attention map. For each sub-node $v\in\mathcal{C}_u$ of u, $\texttt{att}^{\text{dec}}_{u,v}(\boldsymbol{h}_u)$ is defined as:

[1] As the classic settings of graph models, there is also a 'dummy' node in $\mathcal{V}$, used for interpreting the background class. As it does not interact with other semantic human parts (nodes), we omit this node for concept clarity.

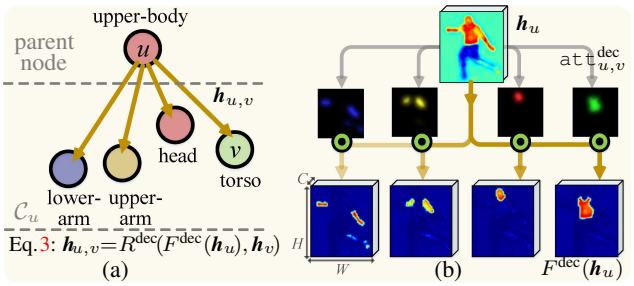


Figure 3: **Illustration of our decompositional relation modeling.** (a) Decompositional relations between the *upper-body* node (u) and its constituents ($\mathcal{C}_u$). (b) With the decompositional attentions $\{\texttt{att}^{\text{dec}}_{u,v}(\boldsymbol{h}_u)\}_{v\in\mathcal{C}_u}$, F^{dec} learns how to 'break down' the *upper-body* node and generates more tractable features for its constituents. In the relation adapted feature $F^{\text{dec}}(\boldsymbol{h}_u)$, the responses from the background and other irrelevant parts are suppressed.

$$\texttt{att}^{\text{dec}}_{u,v}(\boldsymbol{h}_u)=\text{PSM}([\phi^{\text{dec}}_v(\boldsymbol{h}_u)]_{v\in\mathcal{C}_u})=\frac{\exp(\phi^{\text{dec}}_v(\boldsymbol{h}_u))}{\Sigma_{v'\in\mathcal{C}_u}\exp(\phi^{\text{dec}}_{v'}(\boldsymbol{h}_u))}, \quad (4)$$

where PSM$(\cdot)$ stands for *pixel-wise soft-max*, '$[\cdot]$' represents the channel-wise concatenation, and $\phi^{\text{dec}}_v(\boldsymbol{h}_u)\in\mathbb{R}^{W\times H}$ computes a specific significance map for v. 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 over $\boldsymbol{h}_u$. To recap, the *decompositional attention*, conditioned on $\boldsymbol{h}_u$, lets u pass separate high-level information to different child nodes $\mathcal{C}_u$ (see Fig. 3(b)). Here $\texttt{att}^{\text{dec}}_{u,v}(\cdot)$ is node-specific and separately learnt for the three entire nodes in $\mathcal{V}_2\cup\mathcal{V}_3$, namely *full-body*, *upper-body* and *lower-body*. A subscript $_{u,v}$ is added to address this point. In addition, for each parent node u, the groundtruth maps $\mathcal{Y}_{\mathcal{C}_u}=\{y_v\}_{v\in\mathcal{C}_u}\in\{0,1\}^{W\times H\times|\mathcal{C}_u|}$ of all the child nodes $\mathcal{C}_u$ can be used as supervision signals to train its *decompositional attention* $\{\texttt{att}^{\text{dec}}_{u,v}(\boldsymbol{h}_u)\}_{v\in\mathcal{C}_u}\in[0,1]^{W\times H\times|\mathcal{C}_u|}$:

$$\mathcal{L}_{\text{dec}}=\sum\nolimits_{u\in\mathcal{V}_2\cup\mathcal{V}_3}\mathcal{L}_{\text{CE}}\big(\{\texttt{att}^{\text{dec}}_{u,v}(\boldsymbol{h}_u)\}_{v\in\mathcal{C}_u},\mathcal{Y}_{\mathcal{C}_u}\big), \quad (5)$$

where $\mathcal{L}_{\text{CE}}$ represents the standard cross-entropy loss.

2) *Compositional relation modeling:* In the human hierarchy $\mathcal{G}$, compositional relations are represented by vertical, downward edges. To address this type of relations, we design a compositional relation network R^{com} as (Fig. 4):

$$\begin{aligned}\boldsymbol{h}_{u,v}&=R^{\text{com}}(F^{\text{com}}(\boldsymbol{h}_u),\boldsymbol{h}_v),\quad u\in\mathcal{C}_v,\\ F^{\text{com}}(\boldsymbol{h}_u)&=\boldsymbol{h}_u\odot\texttt{att}^{\text{com}}_v([\boldsymbol{h}_{u'}]_{u'\in\mathcal{C}_v}).\end{aligned} \quad (6)$$

Here $\texttt{att}^{\text{com}}_v\colon\mathbb{R}^{W\times H\times c\times|\mathcal{C}_v|}\mapsto[0,1]^{W\times H}$ is a *compositional attention*, implemented by a 1×1 convolutional layer. The rationale behind such a design is that, for a parent node v, $\texttt{att}^{\text{com}}_v$ gathers statistics of all the child nodes $\mathcal{C}_v$ and is used to enhance each sub-node feature $\boldsymbol{h}_u$. As $\texttt{att}^{\text{com}}_v$ is compositional in nature, its enhanced feature $F^{\text{com}}(\boldsymbol{h}_u)$ is more 'friendly' to the parent node v, compared to $\boldsymbol{h}_u$. Thus, R^{com} is able to generate more expressive relation features by considering compositional structures (see Fig. 4(b)).

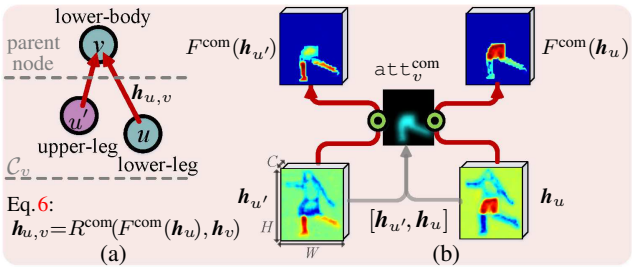


Figure 4: **Illustration of our compositional relation modeling.** (a) Compositional relations between the *lower-body* node (v) and its constituents ($\mathcal{C}_v$). (b) The compositional attention $\texttt{att}^{\text{com}}_v([\boldsymbol{h}_{u'},\boldsymbol{h}_u])$ gathers information from all the constituents $\mathcal{C}_v$ and lets F^{com} enhance all the *lower-body* related features of $\mathcal{C}_v$.

For each parent node $v\in\mathcal{V}_2\cup\mathcal{V}_3$, with its groundtruth map $y_v\in\{0,1\}^{W\times H}$, the *compositional attention* for all its child nodes $\mathcal{C}_v$ is trained by minimizing the following loss:

$$\mathcal{L}_{\text{com}}=\sum\nolimits_{v\in\mathcal{V}_2\cup\mathcal{V}_3}\mathcal{L}_{\text{CE}}\big(\texttt{att}^{\text{com}}_v([\boldsymbol{h}_{u'}]_{u'\in\mathcal{C}_v}),y_v\big). \quad (7)$$

3) *Dependency relation modeling:* In $\mathcal{G}$, dependency relations are represented as horizontal edges (dashed line:-- in Fig. 2(a)), describing pairwise, kinematic connections between human parts, such as (*head*, *torso*), (*upper-leg*, *lower-leg*), *etc*. Two kinematically connected human parts are spatially adjacent, and their dependency relation essentially addresses the context information. For a node u, with its kinematically connected siblings $\mathcal{K}_u$, a dependency relation network R^{dep} is designed as (Fig. 5):

$$\begin{aligned}\boldsymbol{h}_{u,v}&=R^{\text{dep}}(F^{\text{dep}}(\boldsymbol{h}_u),\boldsymbol{h}_v),\quad v\in\mathcal{K}_u,\\ F^{dep}(\boldsymbol{h}_u)&=F^{\text{cont}}(\boldsymbol{h}_u)\odot\texttt{att}^{\text{dep}}_{u,v}\big(F^{\text{cont}}(\boldsymbol{h}_u)\big),\end{aligned} \quad (8)$$

where $F^{\text{cont}}(\boldsymbol{h}_u)\in\mathbb{R}^{W\times H\times c}$ is used to extract the context of u, and $\texttt{att}^{\text{dep}}_{u,v}\big(F^{\text{cont}}(\boldsymbol{h}_u)\big)\in[0,1]^{W\times H}$ is a *dependency attention* that produces an attention for each sibling node v, conditioned on u's context $F^{\text{cont}}(\boldsymbol{h}_u)$. Specifically, inspired by the non-local self-attention [55, 62], the *context extraction* module F^{cont} is designed as:

$$\begin{aligned}F^{\text{cont}}(\boldsymbol{h}_u)&=\rho(\boldsymbol{x}\boldsymbol{A}^\top)\ \in\mathbb{R}^{W\times H\times c},\\ \boldsymbol{A}&=\boldsymbol{h}_u'^{\top}\boldsymbol{W}\boldsymbol{x}'\ \in\mathbb{R}^{(WH)\times(WH)},\end{aligned} \quad (9)$$

where $\boldsymbol{h}'_u\in\mathbb{R}^{(c+8)\times(WH)}$ and $\boldsymbol{x}'\in\mathbb{R}^{(C+8)\times(WH)}$ are node (part) and image representations augmented with spatial information, respectively, flattened into matrix formats. The last eight channels of $\boldsymbol{h}'_u$ and $\boldsymbol{x}'$ encode spatial coordinate information [25], where the first six dimensions are the normalized horizontal and vertical positions, and the last two dimensions are the normalized width and height information of the feature, $1/W$ and $1/H$. $\boldsymbol{W}\in\mathbb{R}^{(c+8)\times(C+8)}$ is learned as a linear transformation based node-to-context projection function. The node feature $\boldsymbol{h}'_u$, used as a *query* term, retrieves the *reference* image feature $\boldsymbol{x}'$ for its context information. As a result, the affinity matrix $\boldsymbol{A}$ stores the attention

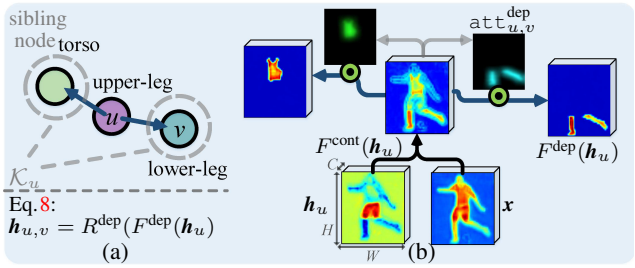


Figure 5: **Illustration of our dependency relation modeling.** (a) Dependency relations between the *upper-body* node (u) and its siblings ($\mathcal{K}_u$). (b) The dependency attention $\{\text{att}^{\text{dep}}_{u,v}(F^{\text{cont}}(\boldsymbol{h}_u))\}_{v\in\mathcal{K}_u}$, derived from u's contextual information $F^{\text{cont}}(\boldsymbol{h}_u)$, gives separate importance for different siblings $\mathcal{K}_u$.

weight between the query and reference at a certain spatial location, accounting for both visual and spatial information. Then, u's context is collected as a weighted sum of the original image feature $\boldsymbol{x}$ with column-wise normalized weight matrix $\boldsymbol{A}^{\top}$: $\boldsymbol{x}\boldsymbol{A}^{\top}\in\mathbb{R}^{C\times(WH)}$. A 1×1 convolution based linear embedding function $\rho:\mathbb{R}^{W\times H\times C}\mapsto\mathbb{R}^{W\times H\times c}$ is applied for feature dimension compression, *i.e.*, to make the channel dimensions of different edge embeddings consistent.

For each sibling node $v\in\mathcal{K}_u$ of u, $\text{att}^{\text{dep}}_{u,v}$ is defined as:

$$\text{att}^{\text{dep}}_{u,v}\big(F^{\text{cont}}(\boldsymbol{h}_u)\big) = \text{PSM}\big([\phi^{\text{dep}}_v(\boldsymbol{h}_u)]_{v\in\mathcal{K}_u}\big). \tag{10}$$

Here $\phi^{\text{dep}}_v(\cdot)\in\mathbb{R}^{W\times H}$ gives an importance map for v, using a 1×1 convolutional layer. Through the *pixel-wise soft-max* operation PSM$(\cdot)$, we enforce $\sum_{v\in\mathcal{K}_u}\text{att}^{\text{dep}}_{u,v}=\mathbf{1}$, leading to a *dependency attention* mechanism which assigns exclusive attentions over $F^{\text{cont}}(\boldsymbol{h}_u)$, for the corresponding sibling nodes $\mathcal{K}_u$. Such a *dependency attention* is learned via:

$$\mathcal{L}_{\text{dep}} = \sum\nolimits_{u\in\mathcal{V}_1\cup\mathcal{V}_2}\mathcal{L}_{\text{CE}}\big(\{\text{att}^{\text{dep}}_{u,v}(\boldsymbol{h}_u)\}_{v\in\mathcal{K}_u},\mathcal{Y}_{\mathcal{K}_u}\big), \tag{11}$$

where $\mathcal{Y}_{\mathcal{K}_u}\in[0,1]^{W\times H\times|\mathcal{K}_u|}$ stands for the groundtruth maps $\{y_v\}_{v\in\mathcal{K}_u}$ of all the sibling nodes $\mathcal{K}_u$ of u.

Iterative inference over human hierarchy: Human bodies present a hierarchical structure. According to graph theory, approximate inference algorithms should be used for such a loopy structure $\mathcal{G}$. However, previous structured human parsers directly produce the final node representation $\boldsymbol{h}_v$ by either simply accounting for the information from the parent node u [74]: $\boldsymbol{h}_v\leftarrow R(\boldsymbol{h}_u,\boldsymbol{h}_v)$, where $v\in\mathcal{C}_u$; or from its neighbors $\mathcal{N}_v$ [61]: $\boldsymbol{h}_v\leftarrow\sum_{u\in\mathcal{N}_v}R(\boldsymbol{h}_u,\boldsymbol{h}_v)$. They ignore the fact that, in such a structured setting, information is organized in a complex system. Iterative algorithms offer a more favorable solution, *i.e.*, the node representation should be updated iteratively by aggregating the messages from its neighbors; after several iterations, the representation can approximate the optimal results [53]. In graph theory parlance, the iterative algorithm can be achieved by a parametric message passing process, which 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 $\mathcal{N}_v$ to enrich the node embedding $\boldsymbol{h}_v$:

$$\begin{aligned}\boldsymbol{m}_v^{(t)} &= \sum\nolimits_{u\in\mathcal{N}_v}M(\boldsymbol{h}_u^{(t-1)},\boldsymbol{h}_v^{(t-1)}),\\ \boldsymbol{h}_v^{(t)} &= U(\boldsymbol{h}_v^{(t-1)},\boldsymbol{m}_v^{(t)}),\end{aligned} \tag{12}$$

where $\boldsymbol{h}_v^{(t)}$ stands for v's state in the t-th iteration. 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 (Fig. 2(e)-(f)):

$$\boldsymbol{m}_v^{(t)} = \underbrace{\sum\nolimits_{u\in\mathcal{P}_v}\boldsymbol{h}_{u,v}^{(t-1)}}_{\text{decomposition}} + \underbrace{\sum\nolimits_{u\in\mathcal{C}_v}\boldsymbol{h}_{u,v}^{(t-1)}}_{\text{composition}} + \underbrace{\sum\nolimits_{u\in\mathcal{K}_v}\boldsymbol{h}_{u,v}^{(t-1)}}_{\text{dependency}}, \tag{13}$$

$$\boldsymbol{h}_v^{(t)} = U_{\text{convGRU}}(\boldsymbol{h}_v^{(t-1)},\boldsymbol{m}_v^{(t)}), \tag{14}$$

where the initial state $\boldsymbol{h}_v^{(0)}$ is obtained by Eq. 1. Here, the message aggregation step (Eq. 13) is achieved by per-edge relation function terms, *i.e.*, node v updates its state $\boldsymbol{h}_v$ by absorbing all the incoming information along different relations. As for the update function U in Eq. 14, we use a convGRU [54], which replaces the fully-connected units in the original MLP-based GRU with convolution operations, to describe its repeated activation behavior and address the pixel-wise nature of human parsing, simultaneously. Compared to previous parsers, which are typically based on *feed-forward* architectures, our massage-passing inference essentially provides a *feed-back* mechanism, encouraging effective reasoning over the cyclic human hierarchy $\mathcal{G}$.

Loss function: In each step t, to obtain the predictions $\hat{\mathcal{Y}}_l^{(t)}=\{\hat{y}_v^{(t)}\in[0,1]^{W\times H}\}_{v\in\mathcal{V}_l}$ of the l-th layer nodes $\mathcal{V}_l$, we apply a convolutional readout function $O:\mathbb{R}^{W\times H\times c}\mapsto\mathbb{R}^{W\times H}$ over $\{\boldsymbol{h}_v^{(t)}\}_{v\in\mathcal{V}}$ (▱ in Fig. 2(g)), and *pixel-wise soft-max* (PSM) for normalization:

$$\hat{\mathcal{Y}}_l^{(t)} = \{\hat{y}_v^{(t)}\}_{v\in\mathcal{V}_l} = \text{PSM}\big([O(\boldsymbol{h}_v^{(t)})]_{v\in\mathcal{V}_l}\big). \tag{15}$$

Given the hierarchical human parsing results $\{\hat{\mathcal{Y}}_l^{(t)}\}_{l=1}^3$ and corresponding groundtruths $\{\mathcal{Y}_l\}_{l=1}^3$, the learning task in the iterative inference can be posed as the minimization of the following loss (Fig. 2(h)):

$$\mathcal{L}_{\text{parsing}}^{(t)} = \sum\nolimits_{l=1}^{3}\mathcal{L}_{\text{CE}}^{(t)}(\hat{\mathcal{Y}}_l^{(t)},\mathcal{Y}_l). \tag{16}$$

Considering Eqs. 5, 7, 11, and 16, the overall loss is defined as:

$$\mathcal{L} = \sum\nolimits_{t=1}^{T}\big(\mathcal{L}_{\text{parsing}}^{(t)} + \alpha(\mathcal{L}_{\text{com}}^{(t)} + \mathcal{L}_{\text{dec}}^{(t)} + \mathcal{L}_{\text{dep}}^{(t)})\big), \tag{17}$$

where the coefficient α is empirically set as 0.1. We set the total inference time $T=2$ and study how the performance changes with the number of inference iterations in §4.3.

3.3. Implementation Details

Node embedding: A DeepLabV3 network [6] serves as the backbone architecture, resulting in a 256-channel image representation whose spatial dimensions are 1/8 of the input image. The projection function $P:\mathbb{R}^{W\times H\times C}\mapsto\mathbb{R}^{W\times H\times c\times|\mathcal{V}|}$ in Eq. 1 is implemented by a 3×3 convolutional layer with

ReLU nonlinearity, where C=256 and $|\mathcal{V}|$ (*i.e.*, the number of nodes) is set according to the settings in different human parsing datasets. We set the channel size of node features c=64 to maintain high computational efficiency.

Relation networks: Each typed relation network R^r in Eq. 2 concatenates the relation-adapted feature $F^r(\boldsymbol{h}_u)$ from the source node u and the destination node v's feature $\boldsymbol{h}_v$ as the input, and outputs the relation representations: $\boldsymbol{h}_{u,v} = R^r([F^r(\boldsymbol{h}_u), \boldsymbol{h}_v])$. $R^r: \mathbb{R}^{W\times H\times 2c} \mapsto \mathbb{R}^{W\times H\times c}$ is implemented by a 3×3 convolutional layer with ReLU nonlinearity.

Iterative inference: In Eq. 14, the update function U_{convGRU} is implemented by a convolutional GRU with 3×3 convolution kernels. The readout function O in Eq. 15 applies a 1×1 convolution operation on the feature-prediction projection. In addition, before sending a node feature $\boldsymbol{h}_v^{(t)}$ into O, we use a light-weight decoder (built using a principle of upsampling the node feature and merging it with the low-level feature of the backbone network) that outputs the segmentation mask with 1/4 the spatial resolution of the input image.

As seen, all the units of our parser are built on convolution operations, leading to spatial information preservation.

4. Experiments

4.1. Experimental Settings

Datasets:[2] Five standard benchmark datasets [22, 64, 44, 31, 45] are used for performance evaluation. LIP [22] contains 50,462 single-person images, which are collected from realistic scenarios and divided into 30,462 images 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 [64] 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 [31] is a 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 and 1,000 images for training, validation, and test, respectively. PPSS [44] is a collection of 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 variations, illumination changes, and occlusions. There are 1,781 and 1,892 images for training and testing, respectively. Fashion Clothing [45] has 4,371 images gathered from Colorful Fashion Parsing [35], Fashionista [68], and Clothing Co-Parsing [69]. 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.

[2] As the datasets provide different human part labels, we make proper modifications of our human hierarchy. For some labels that do not deliver human structures, such as *hat*, *sun-glasses*, we treat them as isolate nodes.

Method	pixAcc.	Mean Acc.	Mean IoU
SegNet [1]	69.04	24.00	18.17
FCN-8s [41]	76.06	36.75	28.29
DeepLabV2 [4]	82.66	51.64	41.64
Attention [5]	83.43	54.39	42.92
†Attention+SSL [22]	84.36	54.94	44.73
DeepLabV3+ [6]	84.09	55.62	44.80
ASN [43]	-	-	45.41
†SSL [22]	-	-	46.19
MMAN [46]	85.24	57.60	46.93
†SS-NAN [72]	87.59	56.03	47.92
HSP-PRI [26]	85.07	60.54	48.16
†MuLA [47]	88.5	60.5	49.3
PSPNet [70]	86.23	61.33	50.56
CE2P [39]	87.37	63.20	53.10
BraidNet [40]	87.60	66.09	54.42
CNIF [61]	88.03	68.80	57.74
Ours	**89.05**	**70.58**	**59.25**

Table 1: **Comparison of pixel accuracy, mean accuracy and mIoU on LIP `val` [22].** † indicates extra pose information used.

Training: ResNet101 [24], pre-trained on ImageNet [52], is used to initialize our DeepLabV3 [6] backbone. The remaining layers are randomly initialized. We train our model on the five aforementioned datasets with their respective training samples, separately. Following the common practice [39, 21, 61], we randomly augment each training sample with a scaling factor in [0.5, 2.0], crop size of 473×473, and horizontal flip. For optimization, we use the standard SGD solver, with a momentum of 0.9 and weight_decay of 0.0005. To schedule the learning rate, we employ the polynomial annealing procedure [4, 70], where the learning rate is multiplied by $(1-\frac{iter}{total\_iter})^{power}$ with *power* as 0.9.

Testing: For each test sample, we set the long side of the image to 473 pixels and maintain the original aspect ratio. As in [70, 47], we average the parsing results over five-scale image pyramids of different scales with flipping, *i.e.*, the scaling factor is 0.5 to 1.5 (with intervals of 0.25).

Reproducibility: Our method is implemented on PyTorch and trained on four NVIDIA Tesla V100 GPUs (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 [72], we report pixel accuracy, mean accuracy and mean Intersection-over-Union (mIoU). For PASCAL-Person-Part and PPSS, following [63, 64, 46], the performance is evaluated in terms of mIoU. For ATR and Fashion Clothing, as in [45, 61], we report pixel accuracy, foreground accuracy, average precision, average recall, and average F1-score.

4.2. Quantitative and Qualitative Results

LIP [22]: LIP is a gold standard benchmark for human parsing. Table 1 reports the comparison results with 16 state-of-

Method	Head	Torso	U-Arm	L-Arm	U-Leg	L-Leg	B.G.	Ave.
HAZN [63]	80.79	59.11	43.05	42.76	38.99	34.46	93.59	56.11
Attention [5]	81.47	59.06	44.15	42.50	38.28	35.62	93.65	56.39
LG-LSTM [33]	82.72	60.99	45.40	47.76	42.33	37.96	88.63	57.97
Attention+SSL [22]	83.26	62.40	47.80	45.58	42.32	39.48	94.68	59.36
Attention+MMAN [46]	82.58	62.83	48.49	47.37	42.80	40.40	94.92	59.91
Graph LSTM [32]	82.69	62.68	46.88	47.71	45.66	40.93	94.59	60.16
SS-NAN [72]	86.43	67.28	51.09	48.07	44.82	42.15	97.23	62.44
Structure LSTM [30]	82.89	67.15	51.42	48.72	51.72	45.91	97.18	63.57
Joint [64]	85.50	67.87	54.72	54.30	48.25	44.76	95.32	64.39
DeepLabV2 [4]	-	-	-	-	-	-	-	64.94
MuLA [47]	84.6	68.3	57.5	54.1	49.6	46.4	95.6	65.1
PCNet [74]	86.81	69.06	55.35	55.27	50.21	48.54	96.07	65.90
Holistic [29]	86.00	69.85	56.63	55.92	51.46	48.82	95.73	66.34
WSHP [15]	87.15	72.28	57.07	56.21	52.43	50.36	**97.72**	67.60
DeepLabV3+ [6]	87.02	72.02	60.37	57.36	53.54	48.52	96.07	67.84
SPGNet [8]	87.67	71.41	61.69	60.35	52.62	48.80	95.98	68.36
PGN [21]	**90.89**	75.12	55.83	64.61	55.42	41.57	95.33	68.40
CNIF [61]	88.02	72.91	64.31	63.52	55.61	54.96	96.02	70.76
Ours	89.73	**75.22**	**66.87**	**66.21**	**58.69**	**58.17**	96.94	**73.12**

Table 2: **Per-class comparison of mIoU on PASCAL-Person-Part `test` [64].**

Method	pixAcc.	F.G. Acc.	Prec.	Recall	F-1
Yamaguchi [68]	84.38	55.59	37.54	51.05	41.80
Paperdoll [67]	88.96	62.18	52.75	49.43	44.76
M-CNN [36]	89.57	73.98	64.56	65.17	62.81
ATR [31]	91.11	71.04	71.69	60.25	64.38
DeepLabV2 [4]	94.42	82.93	78.48	69.24	73.53
PSPNet [70]	95.20	80.23	79.66	73.79	75.84
Attention [5]	95.41	85.71	81.30	73.55	77.23
DeepLabV3+ [6]	95.96	83.04	80.41	78.79	79.49
Co-CNN [34]	96.02	83.57	84.95	77.66	80.14
LG-LSTM [33]	96.18	84.79	84.64	79.43	80.97
TGPNet [45]	96.45	87.91	83.36	80.22	81.76
CNIF [61]	96.26	87.91	84.62	86.41	85.51
Ours	**96.84**	**89.23**	**86.17**	**88.35**	**87.25**

Table 3: **Comparison of accuracy, foreground accuracy, average precision, recall and F1-score on ATR `test` [31].**

the-arts on LIP `val`. We first find that general semantic segmentation methods [1, 41, 4, 6] tend to perform worse than human parsers. This indicates the importance of reasoning human structures in this problem. In addition, though recent human parsers gain impressive results, our model still outperforms all the competitors by a large margin. For instance, in terms of pixAcc., mean Acc., and mean IoU, our parser dramatically surpasses the best performing method, CNIF [61], by 1.02%, 1.78% and 1.51%, respectively. We would also like to mention that our parser does not use additional pose [22, 72, 47] or edge [39] information.

PASCAL-Person-Part [64]: In Table 2, we compare our method against 18 recent methods on PASCAL-Person-Part `test` using IoU score. From the results, we can again see that our approach achieves better performance compared to all other methods; specially, 73.12% *vs* 70.76% of CNIF [61] and 68.40% of PGN [21], in terms of *mIoU*. Such a performance gain is particularly impressive considering that improvement on this dataset is very challenging.

ATR [31]: Table 3 presents comparisons with 14 previous methods on ATR `test`. Our approach sets new state-of-the-arts for all five metrics, outperforming all other methods by a large margin. For example, our parser provides a considerable performance gain in F-1 score, *i.e.*, 1.74% and 5.49% higher than the current top-two performing methods, CNIF [61] and TGPNet [45], respectively.

Method	pixAcc.	F.G. Acc.	Prec.	Recall	F-1
Yamaguchi [68]	81.32	32.24	23.74	23.68	22.67
Paperdoll [67]	87.17	50.59	45.80	34.20	35.13
DeepLabV2 [4]	87.68	56.08	35.35	39.00	37.09
Attention [5]	90.58	64.47	47.11	50.35	48.68
TGPNet [45]	91.25	66.37	50.71	53.18	51.92
CNIF [61]	92.20	68.59	56.84	59.47	58.12
Ours	**93.12**	**70.57**	**58.73**	**61.72**	**60.19**

Table 4: **Comparison of pixel accuracy, foreground pixel accuracy, average precision, average recall and average f1-score on Fashion Clothing `test` [45].**

Method	Head	Face	U-Cloth	Arms	L-Cloth	Legs	B.G.	Ave.
DL [44]	22.0	29.1	57.3	10.6	46.1	12.9	68.6	35.2
DDN [44]	35.5	44.1	68.4	17.0	61.7	23.8	80.0	47.2
ASN [43]	51.7	51.0	65.9	29.5	52.8	20.3	83.8	50.7
MMAN [46]	53.1	50.2	69.0	29.4	55.9	21.4	85.7	52.1
LCPC [9]	55.6	46.6	71.9	30.9	58.8	24.6	86.2	53.5
CNIF [61]	67.6	60.8	80.8	46.8	69.5	28.7	90.6	60.5
Ours	**68.8**	**63.2**	**81.7**	**49.3**	**70.8**	**32.0**	**91.4**	**65.3**

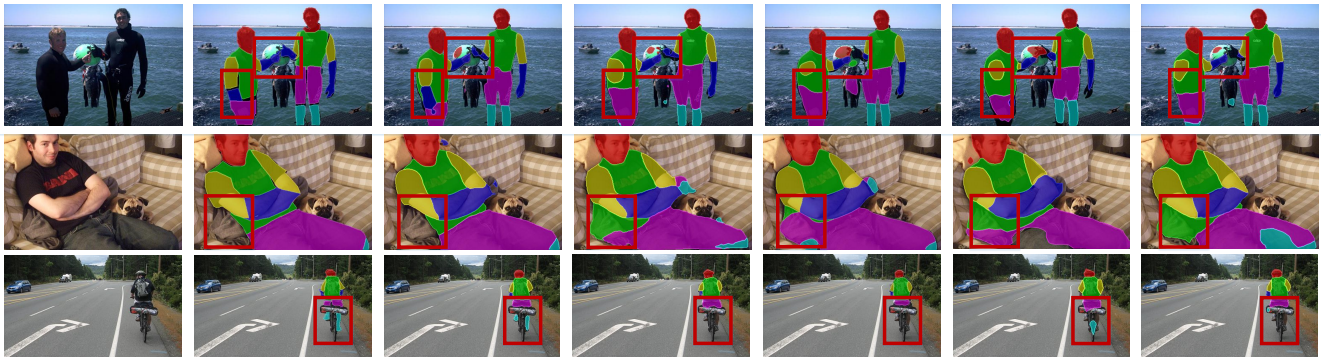
Table 5: **Comparison of mIoU on PPSS `test` [44].**

Fashion Clothing [45]: The quantitative comparison results with six competitors on Fashion Clothing `test` are summarized in Table 4. Our model yields an F-1 score of 60.19%, while those for Attention [5], TGPNet [45], and CNIF [61] are 48.68%, 51.92%, and 58.12%, respectively. This again demonstrates our superior performance.

PPSS [44]: Table 5 compares our method against six famous methods on PPSS `test` set. The evaluation results demonstrate that our human parser achieves 65.3% mIoU, with substantial gains over the second best, CNIF [61], and third best, LCPC [9], of 4.8% and 11.8%, respectively.

Runtime comparison: As our parser does not require extra pre-/post-processing steps (*e.g.*, human pose used in [64], over-segmentation in [32, 30], and CRF in [64]), it achieves a high speed of 12fps (on PASCAL-Person-Part), faster than most of the counterparts, such as Joint [64] (0.1fps), Attention+SSL [22] (2.0fps), MMAN [46] (3.5fps), SS-NAN [72] (2.0fps), and LG-LSTM [33] (3.0fps).

Qualitative results: Some qualitative comparison results on PASCAL-Person-Part `test` are depicted in Fig. 6. We can see that our approach outputs more precise parsing results than other competitors [6, 21, 72, 61], despite the existence of rare pose (2_{nd} row) and occlusion (3_{rd} row). In addition, with its better understanding of human structures, our parser gets more robust results and eliminates the interference from the background (1_{st} row). The last row gives a challenging case, where our parser still correctly recognizes the confusing parts of the person in the middle.



(a) Image (b) Ground-truth (c) Ours (d) DeepLabV3+ [6] (e) PGN [21] (f) SS-NAN [72] (g) CNIF [61]

Figure 6: **Visual comparison on PASCAL-Person-Part `test`.** Our model (c) generates more accurate predictions, compared to other famous methods [6, 21, 72, 61] (d-g). The improved labeled results by our parser are denoted in red boxes. Best viewed in color.

4.3. Diagnostic Experiments

To demonstrate how each component in our parser contributes to the performance, a series of ablation experiments are conducted on PASCAL-Person-Part `test`.

Type-specific relation modeling: We first investigate the necessity of comprehensively exploring different relations, and discuss the effective of our type-specific relation modeling strategy. Concretely, we studied six variant models, as listed in Table 6: **(1)** 'Baseline' denotes the approach only using the initial node embeddings $\{\boldsymbol{h}_v^{(0)}\}_{v\in\mathcal{V}}$ without any relation information; **(2)** 'Type-agnostic' shows the performance when modeling different human part relations in a type-agnostic manner: $\boldsymbol{h}_{u,v}=R([\boldsymbol{h}_u,\boldsymbol{h}_v])$; **(3)** 'Type-specific *w/o* F^r' gives the performance without the relation-adaption operation F^r in Eq. 2: $\boldsymbol{h}_{u,v}=R^r([\boldsymbol{h}_u,\boldsymbol{h}_v])$; **(4-6)** 'Decomposition relation', 'Composition relation' and 'Dependency relation' are three variants that only consider the corresponding single one of the three kinds of relation categories, using our type-specific relation modeling strategy (Eq. 2). Four main conclusions can be drawn: **(1)** Structural information are essential for human parsing, as all the structured models outperforms 'Baseline'. **(2)** Typed relation modeling leads to more effective human structure learning, as 'Type-specific *w/o* F^r' improves 'Type-agnostic' by 1.28%. **(3)** Exploring different kinds of relations are meaningful, as the variants using individual relation types outperform 'Baseline' and our full model considering all the three kinds of relations achieves the best performance. **(4)** Encoding relation-specific constrains helps with relation pattern learning as our full model is better than the one without relation-adaption, 'Type-specific *w/o* F^r'.

Iterative inference: Table 6 shows the performance of our parser with regard to the iteration step t as denoted in Eq. 13 and Eq. 14. Note that, when $t=0$, only the initial node feature is used. It can be observed that setting $T=2$ or $T=3$ provided a consistent boost in accuracy of 4∼5%, on average, compared to $T=0$; however, increasing T beyond 3 gave marginal returns in performance (around 0.1%). Accordingly, we choose $T=2$ for a better trade-off between accuracy and computation time.

Component	Module	mIoU	△	time (ms)
Reference	**Full model** (2 iterations)	73.12	-	81
Relation modeling	Baseline	68.84	-4.28	46
	Type-agnostic	70.37	-2.75	55
	Type-specific *w/o* F^r	71.65	-1.47	55
	Decomposition relation	71.38	-1.74	50
	Composition relation	69.35	-3.77	49
	Dependency relation	69.43	-3.69	52
Iterative Inference T	0 iteration	68.84	-4.28	46
	1 iterations	72.17	-0.95	59
	3 iterations	73.19	+0.07	93
	4 iterations	73.22	+0.10	105
	5 iterations	73.23	+0.11	116

Table 6: **Ablation study (§4.3) on PASCAL-Person-Part `test`.**

5. Conclusion

In the human semantic parsing task, structure modeling is an essential, albeit inherently difficult, avenue to explore. This work proposed a hierarchical human parser that addresses this issue in two aspects. First, three distinct relation networks are designed to precisely describe the compositional/decompositional relations between constituent and entire parts and help with the dependency learning over kinetically connected parts. Second, to address the inference over the loopy human structure, our parser relies on a convolutional, message passing based approximation algorithm, which enjoys the advantages of iterative optimization and spatial information preservation. The above designs enable strong performance across five widely adopted benchmark datasets, at times outperforming all other competitors.

Acknowledgements This work was partially sponsored by Zhejiang Lab's Open Fund (No. 2019KD0AB04), Zhejiang Lab's International Talent Fund for Young Professionals, and CCF-Tencent Open Fund. This work was also partially supported by DARPA XAI grant N66001-17-2-4029, ARO grant W911NF-18-1-0296, the National Key R&D Program of China (Grant Nos. 2018AAA0102800 and 2018AAA0102802) and National Natural Science Foundation of China (Grant No. 61632018).

References

[1] Vijay Badrinarayanan, Alex Kendall, and Roberto Cipolla. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE TPAMI*, 39(12):2481–2495, 2017. 6, 7

[2] Yihang Bo and Charless C Fowlkes. Shape-based pedestrian parsing. In *CVPR*, 2011. 2

[3] Hong Chen, Zi Jian Xu, Zi Qiang Liu, and Song Chun Zhu. Composite templates for cloth modeling and sketching. In *CVPR*, 2006. 2

[4] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE TPAMI*, 40(4):834–848, 2018. 6, 7

[5] Liang-Chieh Chen, Yi Yang, Jiang Wang, Wei Xu, and Alan L Yuille. Attention to scale: Scale-aware semantic image segmentation. In *CVPR*, 2016. 2, 6, 7

[6] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *ECCV*, 2018. 3, 5, 6, 7, 8

[7] Xianjie Chen, Roozbeh Mottaghi, Xiaobai Liu, Sanja Fidler, Raquel Urtasun, and Alan Yuille. Detect what you can: Detecting and representing objects using holistic models and body parts. In *CVPR*, 2014. 2

[8] Bowen Cheng, Liang-Chieh Chen, Yunchao Wei, Yukun Zhu, Zilong Huang, Jinjun Xiong, Thomas Huang, Wen-Mei Hwu, and Honghui Shi. Spgnet: Semantic prediction guidance for scene parsing. In *ICCV*, 2019. 7

[9] Kang Dang and Junsong Yuan. Location constrained pixel classifiers for image parsing with regular spatial layout. In *BMVC*, 2014. 7

[10] Jian Dong, Qiang Chen, Xiaohui Shen, Jianchao Yang, and Shuicheng Yan. Towards unified human parsing and pose estimation. In *CVPR*, 2014. 2

[11] Jian Dong, Qiang Chen, Wei Xia, Zhongyang Huang, and Shuicheng Yan. A deformable mixture parsing model with parselets. In *ICCV*, 2013. 2

[12] David K Duvenaud, Dougal Maclaurin, Jorge Iparraguirre, Rafael Bombarell, Timothy Hirzel, Alán Aspuru-Guzik, and Ryan P Adams. Convolutional networks on graphs for learning molecular fingerprints. In *NIPS*, 2015. 2

[13] S Eslami and Christopher Williams. A generative model for parts-based object segmentation. In *NIPS*, 2012. 2

[14] Lifeng Fan, Wenguan Wang, Siyuan Huang, Xinyu Tang, and Song-Chun Zhu. Understanding human gaze communication by spatio-temporal graph reasoning. In *ICCV*, 2019. 1

[15] Hao-Shu Fang, Guansong Lu, Xiaolin Fang, Jianwen Xie, Yu-Wing Tai, and Cewu Lu. Weakly and semi supervised human body part parsing via pose-guided knowledge transfer. In *CVPR*, 2018. 1, 2, 7

[16] Hao-Shu Fang, Chenxi Wang, Minghao Gou, and Cewu Lu. Graspnet: A large-scale clustered and densely annotated datase for object grasping. *arXiv preprint arXiv:1912.13470*, 2019. 1

[17] Hao-Shu Fang, Shuqin Xie, Yu-Wing Tai, and Cewu Lu. Rmpe: Regional multi-person pose estimation. In *ICCV*, 2017. 2

[18] Hao-Shu Fang, Yuanlu Xu, Wenguan Wang, Xiaobai Liu, and Song-Chun Zhu. Learning pose grammar to encode human body configuration for 3d pose estimation. In *AAAI*, 2018. 2

[19] Justin Gilmer, Samuel S Schoenholz, Patrick F Riley, Oriol Vinyals, and George E Dahl. Neural message passing for quantum chemistry. In *ICML*, 2017. 2

[20] Ke Gong, Yiming Gao, Xiaodan Liang, Xiaohui Shen, Meng Wang, and Liang Lin. Graphonomy: Universal human parsing via graph transfer learning. In *CVPR*, 2019. 1, 2

[21] Ke Gong, Xiaodan Liang, Yicheng Li, Yimin Chen, Ming Yang, and Liang Lin. Instance-level human parsing via part grouping network. In *ECCV*, 2018. 6, 7, 8

[22] Ke Gong, Xiaodan Liang, Dongyu Zhang, Xiaohui Shen, and Liang Lin. Look into person: Self-supervised structure-sensitive learning and a new benchmark for human parsing. In *CVPR*, 2017. 1, 2, 6, 7

[23] William L Hamilton, Rex Ying, and Jure Leskovec. Representation learning on graphs: Methods and applications. *arXiv preprint arXiv:1709.05584*, 2017. 2

[24] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 6

[25] Ronghang Hu, Marcus Rohrbach, and Trevor Darrell. Segmentation from natural language expressions. In *ECCV*, 2016. 4

[26] Mahdi M. Kalayeh, Emrah Basaran, Muhittin Gokmen, Mustafa E. Kamasak, and Mubarak Shah. Human semantic parsing for person re-identification. In *CVPR*, 2018. 6

[27] Thomas N Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. In *ICLR*, 2017. 2

[28] Lubor Ladicky, Philip HS Torr, and Andrew Zisserman. Human pose estimation using a joint pixel-wise and part-wise formulation. In *CVPR*, 2013. 2

[29] Qizhu Li, Anurag Arnab, and Philip HS Torr. Holistic, instance-level human parsing. In *BMVC*, 2017. 7

[30] Xiaodan Liang, Liang Lin, Xiaohui Shen, Jiashi Feng, Shuicheng Yan, and Eric P Xing. Interpretable structure-evolving lstm. In *CVPR*, 2017. 7

[31] Xiaodan Liang, Si Liu, Xiaohui Shen, Jianchao Yang, Luoqi Liu, Jian Dong, Liang Lin, and Shuicheng Yan. Deep human parsing with active template regression. *IEEE TPAMI*, 37(12):2402–2414, 2015. 2, 6, 7

[32] Xiaodan Liang, Xiaohui Shen, Jiashi Feng, Liang Lin, and Shuicheng Yan. Semantic object parsing with graph lstm. In *ECCV*, 2016. 2, 7

[33] Xiaodan Liang, Xiaohui Shen, Donglai Xiang, Jiashi Feng, Liang Lin, and Shuicheng Yan. Semantic object parsing with local-global long short-term memory. In *CVPR*, 2016. 2, 7

[34] Xiaodan Liang, Chunyan Xu, Xiaohui Shen, Jianchao Yang, Si Liu, Jinhui Tang, Liang Lin, and Shuicheng Yan. Human parsing with contextualized convolutional neural network. In *ICCV*, 2015. 2, 7

[35] Si Liu, Jiashi Feng, Csaba Domokos, Hui Xu, Junshi Huang, Zhenzhen Hu, and Shuicheng Yan. Fashion parsing with weak color-category labels. *TMM*, 16(1):253–265, 2014. 2, 6

[36] Si Liu, Xiaodan Liang, Luoqi Liu, Xiaohui Shen, Jian-

chao Yang, Changsheng Xu, Liang Lin, Xiaochun Cao, and Shuicheng Yan. Matching-cnn meets knn: Quasi-parametric human parsing. In *CVPR*, 2015. 2, 7

[37] Si Liu, Yao Sun, Defa Zhu, Guanghui Ren, Yu Chen, Jiashi Feng, and Jizhong Han. Cross-domain human parsing via adversarial feature and label adaptation. In *AAAI*, 2018. 2

[38] Si Liu, Changhu Wang, Ruihe Qian, Han Yu, Renda Bao, and Yao Sun. Surveillance video parsing with single frame supervision. In *CVPR*, 2017. 2

[39] Ting Liu, Tao Ruan, Zilong Huang, Yunchao Wei, Shikui Wei, Yao Zhao, and Thomas Huang. Devil in the details: Towards accurate single and multiple human parsing. *arXiv preprint arXiv:1809.05996*, 2018. 6, 7

[40] Xinchen Liu, Meng Zhang, Wu Liu, Jingkuan Song, and Tao Mei. Braidnet: Braiding semantics and details for accurate human parsing. In *ACMMM*, 2019. 6

[41] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *CVPR*, 2015. 6, 7

[42] Long Zhu, Yuanhao Chen, Yifei Lu, Chenxi Lin, and A. Yuille. Max margin and/or graph learning for parsing the human body. In *CVPR*, 2008. 1, 2

[43] Pauline Luc, Camille Couprie, Soumith Chintala, and Jakob Verbeek. Semantic segmentation using adversarial networks. In *NIPS-workshop*, 2016. 6, 7

[44] Ping Luo, Xiaogang Wang, and Xiaoou Tang. Pedestrian parsing via deep decompositional network. In *ICCV*, 2013. 2, 6, 7

[45] Xianghui Luo, Zhuo Su, Jiaming Guo, Gengwei Zhang, and Xiangjian He. Trusted guidance pyramid network for human parsing. In *ACMMM*, 2018. 2, 6, 7

[46] Yawei Luo, Zhedong Zheng, Liang Zheng, Tao Guan, Junqing Yu, and Yi Yang. Macro-micro adversarial network for human parsing. In *ECCV*, 2018. 2, 6, 7

[47] Xuecheng Nie, Jiashi Feng, and Shuicheng Yan. Mutual learning to adapt for joint human parsing and pose estimation. In *ECCV*, 2018. 1, 2, 6, 7

[48] Mathias Niepert, Mohamed Ahmed, and Konstantin Kutzkov. Learning convolutional neural networks for graphs. In *ICML*, 2016. 2

[49] Seyoung Park, Bruce Xiaohan Nie, and Song-Chun Zhu. Attribute and-or grammar for joint parsing of human pose, parts and attributes. *IEEE TPAMI*, 40(7):1555–1569, 2018. 1

[50] Siyuan Qi, Wenguan Wang, Baoxiong Jia, Jianbing Shen, and Song-Chun Zhu. Learning human-object interactions by graph parsing neural networks. In *ECCV*, 2018. 1

[51] Ingmar Rauschert and Robert T Collins. A generative model for simultaneous estimation of human body shape and pixel-level segmentation. In *ECCV*, 2012. 2

[52] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael S. Bernstein, Alexander C. Berg, and Fei-Fei Li. Imagenet large scale visual recognition challenge. *IJCV*, 115(3):211–252, 2015. 6

[53] Franco Scarselli, Marco Gori, Ah Chung Tsoi, Markus Hagenbuchner, and Gabriele Monfardini. The graph neural network model. *IEEE Transactions on Neural Networks*, 20(1):61–80, 2008. 2, 3, 5

[54] Xingjian Shi, Zhourong Chen, Hao Wang, Dit-Yan Yeung, Wai-Kin Wong, and Wang-chun Woo. Convolutional lstm network: A machine learning approach for precipitation nowcasting. In *NIPS*, 2015. 5

[55] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NIPS*, 2017. 4

[56] Petar Velickovic, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Lio, and Yoshua Bengio. Graph attention networks. In *ICLR*, 2018. 2

[57] Nan Wang and Haizhou Ai. Who blocks who: Simultaneous clothing segmentation for grouping images. In *ICCV*, 2011. 2

[58] Tiancai Wang, Tong Yang, Martin Danelljan, Fahad Shahbaz Khan, Xiangyu Zhang, and Jian Sun. Learning human-object interaction detection using interaction points. In *CVPR*, 2020. 1

[59] Wenguan Wang, Xiankai Lu, Jianbing Shen, David J Crandall, and Ling Shao. Zero-shot video object segmentation via attentive graph neural networks. In *ICCV*, 2019. 2

[60] Wenguan Wang, Yuanlu Xu, Jianbing Shen, and Song-Chun Zhu. Attentive fashion grammar network for fashion landmark detection and clothing category classification. In *CVPR*, 2018. 1

[61] Wenguan Wang, Zhijie Zhang, Siyuan Qi, Jianbing Shen, Yanwei Pang, and Ling Shao. Learning compositional neural information fusion for human parsing. In *ICCV*, 2019. 1, 2, 3, 5, 6, 7, 8

[62] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. In *CVPR*, 2018. 4

[63] Fangting Xia, Peng Wang, Liang-Chieh Chen, and Alan L Yuille. Zoom better to see clearer: Human and object parsing with hierarchical auto-zoom net. In *ECCV*, 2016. 2, 6, 7

[64] Fangting Xia, Peng Wang, Xianjie Chen, and Alan L Yuille. Joint multi-person pose estimation and semantic part segmentation. In *CVPR*, 2017. 1, 2, 6, 7

[65] Fangting Xia, Jun Zhu, Peng Wang, and Alan L Yuille. Pose-guided human parsing by an and/or graph using pose-context features. In *AAAI*, 2016. 1, 2

[66] Wenqiang Xu, Yonglu Li, and Cewu Lu. Srda: Generating instance segmentation annotation via scanning, reasoning and domain adaptation. In *ECCV*, 2018. 2

[67] Kota Yamaguchi, M Hadi Kiapour, and Tamara L Berg. Paper doll parsing: Retrieving similar styles to parse clothing items. In *ICCV*, 2013. 2, 7

[68] Kota Yamaguchi, M Hadi Kiapour, Luis E Ortiz, and Tamara L Berg. Parsing clothing in fashion photographs. In *CVPR*, 2012. 2, 6, 7

[69] Wei Yang, Ping Luo, and Liang Lin. Clothing co-parsing by joint image segmentation and labeling. In *CVPR*, 2014. 2, 6

[70] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. In *CVPR*, 2017. 6, 7

[71] Jian Zhao, Jianshu Li, Yu Cheng, Terence Sim, Shuicheng Yan, and Jiashi Feng. Understanding humans in crowded scenes: Deep nested adversarial learning and a new benchmark for multi-human parsing. In *ACMMM*, 2018. 2

[72] Jian Zhao, Jianshu Li, Xuecheng Nie, Fang Zhao, Yunpeng Chen, Zhecan Wang, Jiashi Feng, and Shuicheng Yan. Self-supervised neural aggregation networks for human parsing. In *CVPR-workshop*, 2017. 2, 6, 7, 8

[73] Zilong Zheng, Wenguan Wang, Siyuan Qi, and Song-Chun

Zhu. Reasoning visual dialogs with structural and partial observations. In *CVPR*, 2019. 2

[74] Bingke Zhu, Yingying Chen, Ming Tang, and Jinqiao Wang. Progressive cognitive human parsing. In *AAAI*, 2018. 1, 2, 3, 5, 7