Hierarchical Human Semantic Parsing with Comprehensive Part-Relation Modeling

Wenguan Wang, *Member IEEE*, Tianfei Zhou, Siyuan Qi,
Jianbing Shen, *Senior Member IEEE*, Song-Chun Zhu, *Fellow IEEE*

Abstract—Modeling the human structure is central for human parsing that extracts pixel-wise semantic information from images. We start with analyzing three types of inference processes over the hierarchical structure of human bodies: direct inference (directly predicting human semantic parts using image information), bottom-up inference (assembling knowledge from constituent parts), and top-down inference (leveraging context from parent nodes). We then formulate the problem as a compositional neural information fusion (CNIF) framework, which assembles the information from the three inference processes in a conditional manner, *i.e.*, considering the confidence of the sources. Based on CNIF, we further present a part-relation-aware human parser (PRHP), which precisely describes three kinds of human part relations, *i.e.*, decomposition, composition, and dependency, by three distinct relation networks. Expressive relation information can be captured by imposing the parameters in the relation networks to satisfy specific geometric characteristics of different relations. By assimilating generic message-passing networks with their edge-typed, convolutional counterparts, PRHP performs iterative reasoning over the human body hierarchy. With these efforts, PRHP provides a more general and powerful form of CNIF, and lays the foundation for more sophisticated and flexible human relation patterns of reasoning. Experiments on five datasets demonstrate that our two human parsers outperform the state-of-the-arts in all cases.

Index Terms—Human Parsing, Hierarchical Model, Relation Reasoning, Graph Neural Network.

1 Introduction

Human parsing, which segments human bodies into semantic parts (*e.g.*, arms, legs, *etc.*), is a crucial yet challenging task for fine-grained human body configuration analysis in 2D monocular images. It has attracted tremendous attention in the literature, as it finds a wide spectrum of human-centric applications, such as surveillance analysis, and human-robot interaction, *etc*.

Human bodies present a highly structured hierarchy and body parts inherently interact with each other. Thus the central problem in human parsing is how to model the structures. Though recent human parsers have made remarkable progress, such problem is far from solved. Specifically, some representative ones built upon well-designed deep learning architectures for semantic segmentation (*e.g.*, fully convolutional networks (FCNs) [3], DeepLab [4], *etc.*), failing to utilize the rich structures in this task. Some others only leverage extra human joints to constrain body configurations [5]–[8], causing them suffer from trivial structural information, not to mention the need of extra pose annotations. In this paper, we explore a third direction: to exploit the hierarchical nature of the human body structure as shown in Fig. 1(a-c). Here we solve a slightly augmented problem: besides only segmenting the fine-grained semantic parts (leaf nodes in the human structural hierarchy), we find the segmentation of body parts of all levels. This allows us to make full use of human part-relations and enables a more comprehensive human semantic understanding.

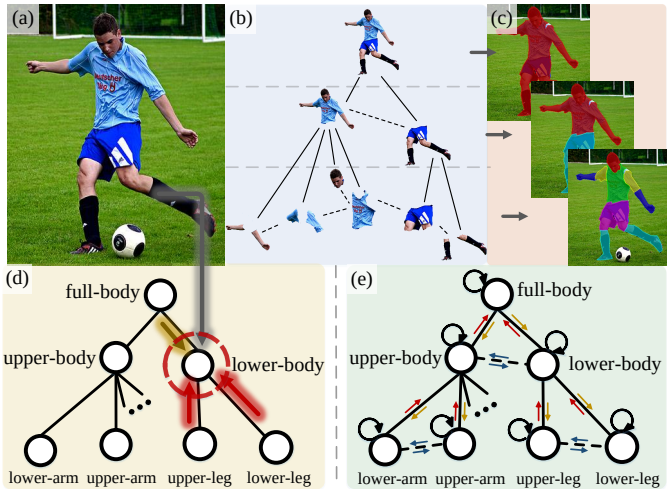


Fig. 1. With the hierarchical human body representation (b), we explore structures for complete human semantic understanding (c). Here - - and / indicate dependency and (de-)compositional relations, respectively. We first propose a **compositional neural information fusion** (CNIF) based parser (d), which fuses information from three sources, *i.e.*, direct, bottom-up, and top-down processes, to infer each part. For clarity, we only show the information fusion of *lower-body* node. We further develop a more powerful **part-relation aware human parser** (PRHP) (e), which isequipped with three distinct relation networks (→, →, and →) to address specific constraints of different part relations (*i.e.*, decomposition, composition, and dependency). In addition, iterative inference (↻) is performed for better approximation.

We first formulate the task as a **multi-source informa-**

- *W. Wang and T. Zhou are with ETH Zurich, Switzerland. (Email: {wenguanwang.ai, ztfei.debug}@gmail.com)*
- *S. Qi is with Google, USA. (Email: syqi@cs.ucla.edu)*
- *J. Shen is with the Inception Institute of Artificial Intelligence, UAE. (Email: shenjianbingcg@gmail.com)*
- *S.-C. Zhu is with Tsinghua University and Peking University, China.*
- *This work builds upon two earlier conference papers, appeared in ICCV2019 [1] and CVPR2020 [2], respectively.*
- *Corresponding author: Jianbing Shen*

tion fusion procedure, based on the insight that the cross-level information within the human body hierarchy can assist learning and inference for each body part. This is also evidenced by human perception studies [9], [10]; a global shape can either precede or follow the recognition of its local parts, and both contribute to the final recognition. In particular, as shown in Fig.1(d), we integrate information from the following three processes for hierarchical human parsing. **1) Direct inference** (or unconscious inference) from the input image. For example, sometimes humans directly recognize objects relying on intuitive understanding [11], [12]. **2) Top-down inference** (), which recognizes fine-grained components from a whole entity. For example, when recognizing small fine-grained parts, exploring contextual information of the entire object is essential [13], [14]. **3) Bottom-up inference** (), which associates constituent parts to predict upper-level nodes. When objects are partially occluded or contain complex topologies, humans can assemble sub-parts to assist in recognizing the entities [15]. We build a **compositional neural information fusion** (CNIF) framework for these three inference processes in an end-to-end manner, yielding a hierarchical human parser that better captures the compositional constraints and structured semantics. In addition, we design CNIF as a conditional fusion, *i.e.*, the assembly of different information is dependent on the confidence estimations for the sources, instead of simply assuming all the sources are reliable. This is achieved by a learnable gate mechanism, leading to more accurate parsing results.

With above efforts, we further consider a more complete set of human part relations. As shown in Fig. 1(b), we address three different relations between human parts [16], [17]: **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*). In CNIF, the bottom-up and top-down inferences are essentially to model the compositional and decompositional relations, respectively. To enable a deeper understanding of human structures, we further develop a **part-relation aware human parser** (PRHP), which precisely and completely describes diverse part relations, and efficiently reasons structures with the prism of a message-passing, feed-back inference scheme (Fig. 1(e)). Specifically, three distinct relation networks (, , and) are designed to explicitly satisfy the specific, intrinsic properties of the three different part-relations, namely decomposition, composition, and dependency. As the human body yields a complex, directed, and cyclic inference graph, it is desirable to run iterative inference for optimal result approximation. To address this issue, a modified, relation-typed convolutional message passing procedure () is performed over the human body hierarchy, enabling our method to obtain better parsing results from a global view. All the components, *i.e.*, the part nodes, edge (relation) functions, and message passing modules, are fully differentiable, making PRHP end-to-end trainable and, in turn, facilitating learning about parts, relations, and inference algorithms. PRHP yields a more general and powerful form of CNIF and is favored in complete and precise part-relation modeling as well as iterative inference for optimal result approximation.

Our main **contributions** are summarized as follows:

1) Through representing human bodies as a hierarchy of multi-level semantic parts, we integrate the strong learning capability of neural networks and the powerful representation ability of graph models, to efficiently mine the inherent structures and capture human semantics from a comprehensive view (§3.1).
2) We first propose a compositional neural information fusion (CNIF) based human parsing network, which tackles the problem as a neural information fusion process over the human body hierarchy (§3.2). It end-to-end incorporates the information from direct, top-down, and bottom-up inference processes while considering the reliability of each. Thus it is able to capture the compositional relations within human structures and enforce high-level constraints over parsing results.
3) We further propose a part-relation aware human parser (PRHP), which utilizes three distinct relation networks to address specific intrinsic properties of different part relations (*i.e.*, decomposition, composition, and dependency). In addition, relation-typed, convolutional message passing is performed over the loopy human body hierarchy for effective and iterative reasoning (§3.3).

We evaluate our two models on five standard human parsing datasets (*i.e.*, LIP [6], PASCAL-Person-Part [7], ATR [18], Fashion Clothing [19], and PPSS [19]). Testing with more than 20K images demonstrates the superiority over existing methods of exploiting structural information within human bodies (§4.2). In addition, with ablation studies for each essential component in our parsers (§4.5), four key insights are found: **1)** Modeling human semantics from a hierarchical view indeed boosts the performance. **2)** Exploring structures is valuable for human parsing. **3)** Distinctly and explicitly modeling different types of relations can better support human structure reasoning. **4)** Message passing based feed-back inference is able to reinforce parsing results.

The present work builds upon our two earlier conference papers [1], [2]. In [1], we proposed to tackle hierarchical human parsing through fusing information from the direct, top-down, and bottom-up inference processes. In [2], we learned the compositional, decompositional, and dependency relations in a type-wise manner and addressed iterative reasoning. For the present paper we have consolidated the overall technique. Moreover, through an ablation study on multiple variants derived from the algorithm, we quantitatively demonstrate the effectiveness of our main points. Last but not least, we report more extensive experimental comparisons with recent methods, for further validation. Our implementations have been made publicly available[1].

2 Related Work

Hierarchical/Graph Models in Computer Vision. Hierarchical/graph models are powerful for building structured representations, which can reflect task-specific relations and constraints. From early distributional semantic models [20], [21], part-based models [22], [23], MRF/CRF [24], And-Or grammar model [25], to deep structural networks [26], [27],

1. CNIF: https://github.com/ZzzjzzZ/CompositionalHumanParsing
PRHP: https://github.com/hlzhu09/Hierarchical-Human-Parsing

hierarchical networks [14], [28], graph neural networks [29], trainable CRF [30], *etc*., hierarchical/graph models have found applications in a wide variety of core computer vision tasks, such as object recognition [31], human parsing [32]–[34], pose estimation [17], [35], visual dialog [36], *etc*., to the extent that they are now ubiquitous in the field. Inspired by their general success, we augment hierarchical human semantic representations with the learning capability of neural networks. Specifically, in addition to directly inferring segments from the image features, our CNIF further derives two extra inference processes, *i.e*., bottom-up and top-down inference, to better capture human structures. For PRHP, compositional, decompositional, and dependency relations are distinctively modeled and encoded into a heterogeneous graph model to encourage more reasonable results that are consistent with human body configurations.

Information Fusion. Our CNIF model is inspired by the idea of fusing information from different sources to obtain a better prediction of the target. One typical application of this is sensor fusion, which is a broad field that we refer the readers to [37] for a thorough treatment. Many machine learning models can be regarded as information fusion methods: *e.g*., product of experts [38], Bayesian fusion, ensemble methods [39], and graphical models [40]. Motivated by this general idea, CNIF learns to adaptively fuse the direct inference along with top-down and bottom-up information for structured semantic reasoning.

Graph Neural Networks (GNNs). As a part of the huge family of graph learning, GNNs have a rich history (dating back to [41]) and became a veritable explosion in research community over the last few years [42]. GNNs effectively learn graph representations in an end-to-end manner, and can be divided into two broad classes: Graph Convolutional Networks (GCNs) and Message Passing Graph Networks (MPGNs). The former [43]–[45] directly extends classical CNNs to structured, non-Euclidean data. Their simple architecture promotes their popularity, while limits their modeling capability for complex structures [42]. MPGNs [29], [46], [47] parameterize all the nodes, edges, and information fusion steps in graph learning, leading to more complicated yet flexible architectures.

Our PRHP falls in the second category, representing an early attempt that explores GNNs in the area of human parsing and crucially differentiates itself in two aspects. **1)** Most previous MPGNs are edge-type-agnostic (based on homogeneous graphs), while PRHP addresses relation-typed structure reasoning (over heterogeneous graphs) with a higher expressive ability. **2)** By replacing the Multilayer Perceptron (MLP) based MPGN units with convolutional counterparts, PRHP gains a spatial information preserving property, which is desirable for such a pixel-wise prediction task.

Human Semantic Parsing. Over the past decade, active research has been devoted towards pixel-level human semantic understanding. This is because human semantic parsing can benefit a wide range of human-related applications, such as human communication behavior analysis [48]–[50], human-object interaction understanding [31], [51], human pose estimation [27], [52], to name a couple. Early approaches tended to leverage low-level image decompositions (*e.g*., super-pixel) [53]–[55], hand-crafted features [56], [57], part templates [58]–[60] and human keypoints [53]–[55], [61], and typically explored certain heuristics over human body configurations [59], [60], [62] in by CRFs [61], [63], structured models [54], [59], grammar models [16], [60], [62], or generative models [64], [65]. Though achieving impressive results, these pioneering works require a lot of hand-designed pipelines, and suffer from the limited representability of hand-crafted features.

With the renaissance of connectionism in the computer vision community, recent research efforts take deep neural networks as their main building blocks. Some pioneering efforts revisit classic template matching strategy [18], [66], address local and global cues [67], or use tree-LSTMs to gather structure information [32], [33]. However, due to the use of superpixel [32], [33], [67] or HOG feature [68], they are fragmentary and time-consuming. Consequent attempts thus follow a more elegant FCN architecture, addressing multi-level cues [69], [70], feature aggregation [19], [71], [72], adversarial learning [73]–[75], or cross-domain knowledge [75], [76]. To further explore inherent structures, numerous approaches [5]–[8], [71], [77] choose to straightforward encode pose information [78], [79] into the parsers, however, relying on off-the-shelf pose estimators or additional annotations. Rather than these approaches addressing category-level understanding of human semantics, a few recent human parsers are specifically designed for the instance-aware setting [80]–[83].

The aforementioned deep human parsers generally achieve promising results, due to the strong learning power of neural networks [3], [4] and the availability of plentiful annotated data [6], [7]. However, they typically need to pre-segment images into superpixels [32], [33], which breaks the end-to-end story and is time-consuming, or rely on extra human landmarks [5]–[8], [77], requiring additional annotations or pre-trained pose estimators. In contrast, we elaborately design a compositional neural information fusion framework, CNIF, which explicitly captures human compositional structures and dynamically combines direct, bottom-up and top-down inference modes over the hierarchy. The overall model inherits the complementary advantages of FCNs and hierarchical models, yielding a unified, end-to-end trainable human parsing framework with a strong learning ability, improved representational power, as well as high processing speed. Though [34] also performs multi-level, fine-grained parsing, it neither explores different information flows within human body hierarchies nor models the problem from the view of structured learning. In addition, prior efforts largely ignore iterative inference and seldom address explicit relation modeling, easily suffering from weak expressive ability and risk of sub-optimal results. To address these limitations, our PRHP model more precisely models the different relations residing on human bodies, *i.e*., decomposition, composition, and dependency, and addresses iterative, spatial-information preserving inference over the human body hierarchy.

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 shown in Fig. 3(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 semantic parts typically considered in prior human parsers: *head, arm, hand, etc.*), two middle-level nodes $\mathcal{V}_2 = \{$*upper-body, lower-body*$\}$ and one root $\mathcal{V}_3 = \{$*full-body*$\}$[2]. The edge set $\mathcal{E} \subseteq \binom{\mathcal{V}}{2}$ represents the relations between human parts (nodes), *i.e.*, the directed edge $e = (u, v) \in \mathcal{E}$ links node u to $v : u \to v$. Each node v and each edge (u, v) are associated with feature vectors: $\boldsymbol{h}_v$ and $\boldsymbol{h}_{u,v}$, respectively. For each node v, we want to infer a segmentation map $y_v \in \mathcal{Y}$ that is a probability map of its label. The groundtruth maps $\mathcal{Y}$ are also organized in a hierarchical manner: $\mathcal{Y} = \cup_{l=1}^{3} \mathcal{Y}_l$. Please note that such a problem setting does not introduce any additional annotation requirement, since higher-level annotations can be obtained by simply combining the lower-level labels.

3.2 Compositional Neural Information Fusion (CNIF) for Hierarchical Human Parsing

We formulate the task as a neural information fusion procedure. Specifically, for each node v, the prediction of y_v is computed by fusing information from three different sources: 1) the raw input image, 2) y_u for the parent node u, and 3) $\boldsymbol{y_w}$ for all the child nodes $\boldsymbol{w}$. Next, we briefly review different methods for information fusion modeling that motivate our solution for human parsing.

3.2.1 Information Fusion

Information fusion refers to the process of combining information from several sources $Z = \{z_1, z_2, \cdots, z_n\}$ in order to form a unified picture of the measured/predicted target y. Each source provides an estimation of the target. These sources can be the raw data x or some other quantities that can be inferred from x. Several approaches have been proposed to tackle this problem.

• Product of experts (PoE) [38] treats each source as an "expert". It multiplies the probabilities and then renormalizes:

$$p(y|Z) = \frac{\prod_{i=1}^{n} p(y|z_i)}{\sum_y \prod_{i=1}^{n} p(y|z_i)}. \tag{1}$$

• Bayesian fusion. Denoting $Z_s = \{z_1, z_2, \cdots, z_s\}$ as the set of the first s sources, it factorizes the posterior probability:

$$p(y|Z) = \frac{p(Z_n|y)p(y)}{p(Z_n)} = \frac{p(y)p(z_1|y)\prod_{s=2}^{n} p(Z_s|Z_{s-1}, y)}{p(z_1)\prod_{s=2}^{n} p(Z_s|Z_{s-1})}. \tag{2}$$

However, it is too difficult to learn all the conditional distributions. By assuming the independence of different information sources, we have the Naive Bayes:

$$p(y|Z) \propto p(y) \prod\nolimits_i p(z_i|y), \tag{3}$$

which serves as an approximation of the true distribution.

• Ensemble methods. In this approach, each z_i is a classifier that predicts y. A typical ensemble method is Bayesian voting [39], which weights the prediction of each classifier to get the final prediction:

$$p(y|Z) = \sum\nolimits_{z_i} p(y|z_i)p(z_i|x). \tag{4}$$

The AdaBoost [84] algorithm also falls into this category.

2. 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.

• Graphical models (*e.g.*, conditional random fields). In such models, each z_i can be viewed as a node that contributes to the conditional probability:

$$p_\theta(y|Z) = \exp\{\sum\nolimits_i \phi_{\theta_i}(y, z_i) - A(\theta)\}, \tag{5}$$

where $A(\theta)$ is the log-partition function that normalizes the distribution. Computing $A(\theta)$ is often intractable, hence the solution is usually given by approximation methods, such as Monte Carlo methods or (loopy) belief propagation [85].

3.2.2 Compositional Neural Information Fusion

The above methods can all be viewed as ways to approximate the true underlying distribution $p(y|Z)$, which can be written as a function of predictions from different information sources Z:

$$p(y|Z) = f(p(y|z_1), p(y|z_2), \cdots, p(y|z_n)). \tag{6}$$

There are potential drawbacks to following the exact solution of one of the above methods. First, they are not entirely consistent with each other. For example, the PoE multiplies all $p(y|z_i)$ together, whereas ensemble methods compute their weighted sum. Each method approximates the true distribution in a different way and has its own tradeoff. Second, exact inference is difficult and solutions are often approximative (*e.g.*, contrastive divergence [86] is used for PoE and Monte Carlo methods for graphical models).

Therefore, instead of exactly following the computation of one of the above methods, we leverage neural networks to directly model this fusion function, due to their strong ability for flexible feature learning and function approximation [87], [88]. The hope is that we can directly learn to fuse multi-source information for a specific task. However, the fusion network should not be learned arbitrarily without inductive biases [89], which is the preference for structural explanations exhibited in human reasoning processes. Here, we exploit the compositional nature of the problem and design the network with the following observations:

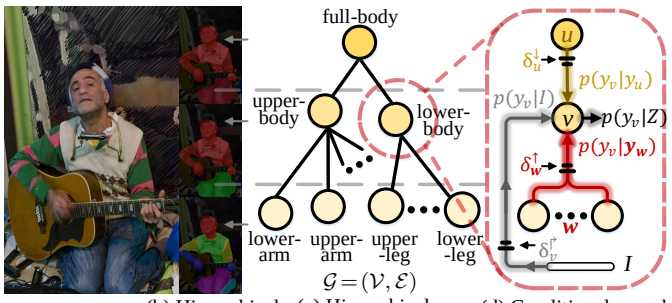
• In the compositional structure $\mathcal{G}$, the final prediction $p(y_v|Z)$ for each node v combines information from three different sources: 1) the direct inference $p(y_v|x)$ from the raw image input, 2) the top-down inference $p(y_v|y_u)$ from the parent node u, which utilizes the **decompositional** relation, and 3) the bottom-up inference $p(y_v|\boldsymbol{y_w})$, which assembles predictions $\boldsymbol{y_w}$ for all the child nodes $\boldsymbol{w}$ to leverage the **compositional** relation.

• In many cases, simply fusing different estimations could be problematic. The final decision should be conditioned on the **confidence** of each information source.

Based on the above observations, we design our parser network to learn a *compositional neural information fusion*:

$$p(y_v|Z) = f(\,\delta_v^{\uparrow} p(y_v|x),\, \delta_u^{\downarrow} p(y_v|y_u),\, \delta_{\boldsymbol{w}}^{\uparrow} p(y_v|\boldsymbol{y_w})\,), \tag{7}$$

where the confidence δ is a learnable continuous function with outputs from 0 to 1. The symbols $\uparrow$, $\downarrow$, and $\uparrow$ denote direct, top-down, and bottom-up inference, respectively. As shown in Fig. 2(d), this function fuses information from the three sources in the compositional structure, taking into account the confidence of each source. For neural network realizations of this function, the probability terms can be relaxed to logits, which are essentially log-probabilities.



(a) Image I (b) Hierarchical human parsing (c) Hierarchical human graph (d) Conditional neural information fusion

Fig. 2. Given an input image (a), our CNIF based human parser performs compositional and conditional neural information fusion over the human semantic graph (c) for hierarchical parsing (b). See §3.2.2 for details.

When carrying out such a prediction, there is one computational issue. Notice that the top-down/bottom-up inference relies on an estimation of the parent/child node(s). This forms a circular dependency between a parent and its children. To solve this, we treat the direct inference result from the raw data as an initial estimation, upon which we perform the top-down/bottom-up inference[3]. Therefore, we decompose the algorithm into three consecutive steps:

1. Direct inference. Given the raw data as input, we assign an estimation $\hat{y}_v$ for each node $v \in \mathcal{V}$.

2. Top-down/bottom-up inference. We estimate $p(y_v|\hat{y}_u)$ and $p(y_v|\hat{\boldsymbol{y}}_{\boldsymbol{w}}))$ based on $\hat{y}_u$ and $\hat{\boldsymbol{y}}_{\boldsymbol{w}}$ estimated in step 1.

3. Conditional information fusion. Based on the above results, we obtain a final prediction for each node v by $y_v^* = \text{argmax}_y\, f(\delta_v^{\uparrow} p(y_v|x),\, \delta_u^{\downarrow} p(y_v|\hat{y}_u),\, \delta_{\boldsymbol{w}}^{\uparrow} p(y_v|\hat{\boldsymbol{y}}_{\boldsymbol{w}}))$.

This procedure motivates the design of our CNIF based human parser, where each step above can be learned as a module by a neural network.

3.2.3 Network Architecture

Our model stacks the following parts to form an end-to-end system for hierarchical human parsing. The system does not require any preprocessing and the modules are fully convolutional, so it is highly efficient.

Direct Inference Network. This directly predicts a segmentation map $\hat{y}_v$ for each node v (a human part), using information from the image (see Fig. 3(b)). Formally, given an input image I, a backbone network F^{B} (*i.e.*, a DeepLabV3-like network, parameterized by $\boldsymbol{W}^{\text{B}}$) is first employed to obtain an image representation $\boldsymbol{h}_I$ (*i.e.*, a (W, H, C)-dimensional tenor encodes full spatial details):

$$\text{image embedding:} \quad \boldsymbol{h}_I = F^{\text{B}}(I; \boldsymbol{W}^{\text{B}}) \in \mathbb{R}^{W\times H\times C}. \tag{8}$$

As the nodes $\mathcal{V}$ capture explicit semantics, a specific feature $\boldsymbol{h}_v$ for each node v is desired for more efficient representation. However, using several different, node-specific embedding networks will lead to a high computational cost. To remedy this, for each l-th level, we first apply a *level-specific* FCN (LSF) to describe the level-wise semantics and contextual relations:

$$\text{level-specific embedding:} \ \boldsymbol{h}_l^{\text{LSF}} = F_l^{\text{LSF}}(\boldsymbol{h}_I; \boldsymbol{W}_l^{\text{LSF}}) \in \mathbb{R}^{W\times H\times c}, \tag{9}$$

where $l \in \{1, 2, 3\}$. Specifically, three LSFs (F_1^{LSF}, F_2^{LSF}, and F_3^{LSF}) are learned to extract three level-specific embeddings ($\boldsymbol{h}_1^{\text{LSF}}$, $\boldsymbol{h}_2^{\text{LSF}}$, and $\boldsymbol{h}_3^{\text{LSF}}$). Further, for each node v, an independent channel-attention block, Squeeze-and-Excitation (SE) [90], is applied to obtain its specific feature:

$$\text{node-specific embedding:} \ \boldsymbol{h}_v = F_v^{\text{SE}}(\boldsymbol{h}_l^{\text{LSF}}; \boldsymbol{W}_v^{\text{SE}}) \in \mathbb{R}^{W\times H\times c}, \tag{10}$$

where $v \in \mathcal{V}_l$ (*i.e.*, v is located in the l-th level). By explicitly modelling the interdependencies between channels, F_v^{SE} allows us to adaptively recalibrate the channel-wise features of $\boldsymbol{h}_l^{\text{LSF}}$ to generate node-wise representations. Meanwhile, due to its light-weight nature, we can achieve our goal with minimal computational overhead. Then, the direct inference network $F^{\uparrow}$ reads the feature and predicts the segmentation map $\hat{y}_v$:

$$\text{logit}(\hat{y}_v|I) = F^{\uparrow}(\boldsymbol{h}_v; \boldsymbol{W}^{\uparrow}) \in \mathbb{R}_{\geq 0}^{W\times H}. \tag{11}$$

Top-down Inference Network. Based on the outputs from the direct inference network, the top-down inference predicts segmentation maps by considering human decompositional structures. Specifically, for node v, the top-down network $F^{\downarrow}$ leverages the initial estimation $\hat{y}_u$ of its parent node u as contextual information for prediction (Fig. 3(c)):

$$\text{logit}(y_v|\hat{y}_u) = F^{\downarrow}(y_v|\hat{y}_u; \boldsymbol{h}_v, \boldsymbol{W}^{\downarrow}) = F^{\downarrow}([\hat{y}_u, \boldsymbol{h}_v]) \in \mathbb{R}_{\geq 0}^{W\times H}. \tag{12}$$

Here, the concatenated feature $[\hat{y}_u, \boldsymbol{h}_v]$ is fed into the FCN-based $F^{\downarrow}$, parameterized by $\boldsymbol{W}^{\downarrow}$, for top-down inference.

Bottom-up Inference Network. One major difference to the top-down network is that, for each node v, the bottom-up network needs to gather information (*i.e.*, $\hat{\boldsymbol{y}}_{\boldsymbol{w}} \in \mathbb{R}_{\geq 0}^{W\times H\times|\boldsymbol{w}|}$) from multiple descendants $\boldsymbol{w}$. Thanks to the compositional relations between $\boldsymbol{w}$ and v, we can transform $\hat{\boldsymbol{y}}_{\boldsymbol{w}}$ to a fixed one-channel representation $\hat{y}_{\boldsymbol{w}}$ through *position-wise max-pooling* PMP (across channels):

$$\hat{y}_{\boldsymbol{w}} = \text{PMP}([\hat{y}_w]_{w\in\boldsymbol{w}}) \in \mathbb{R}_{\geq 0}^{W\times H}, \tag{13}$$

where $[\cdot]$ is a concatenation operation. Then, the bottom-up network $F^{\uparrow}$ gives a prediction according to compositional relations (see Fig. 3(d)):

$$\text{logit}(y_v|\hat{\boldsymbol{y}}_{\boldsymbol{w}}) = F^{\uparrow}(y_v|\hat{\boldsymbol{y}}_{\boldsymbol{w}}; \boldsymbol{h}_v, \boldsymbol{W}^{\uparrow}) = F^{\uparrow}([\hat{y}_{\boldsymbol{w}}, \boldsymbol{h}_v]) \in \mathbb{R}_{\geq 0}^{W\times H}. \tag{14}$$

Conditional Fusion Network. Before making the final prediction, we estimate the confidence δ of each information source using a neural gate function. For the direct inference of a node v, we estimate the confidence by:

$$\delta_v^{\uparrow} = \sigma(\boldsymbol{C}_v^{\uparrow} \cdot \text{CAP}(\boldsymbol{h}_v)) \in [0, 1], \tag{15}$$

where σ is the *sigmoid* function. Here, CAP stands for *channel-wise average pooling*, which has been proved a simple yet effective way for capturing the global statistics of convolutional features [90], [91]. $\boldsymbol{C}_v^{\uparrow} \in \mathbb{R}^{1\times c}$ indicates the parameters of a small fully connected layer which maps the c-dimensional statistic vector $\text{CAP}(\boldsymbol{h}_v) \in \mathbb{R}^c$ of $\boldsymbol{h}_v$ into a confidence score.

The confidence scores for the top-down and bottom-up processes follow a similar computational framework:

$$\begin{aligned} \delta_u^{\downarrow} &= \sigma(\boldsymbol{C}_u^{\downarrow} \cdot \text{CAP}(\boldsymbol{h}_u)) \in [0, 1], \\ \delta_{\boldsymbol{w}}^{\uparrow} &= \sigma(\boldsymbol{C}_{\boldsymbol{w}}^{\uparrow} \cdot \text{CAP}([\boldsymbol{h}_w]_{w\in\boldsymbol{w}})) \in [0, 1], \end{aligned} \tag{16}$$

3. For some nodes, bottom-up or top-down inference might not exist. The terminal leaf nodes $\mathcal{V}_1$ do not have bottom-up inference, while the root node $\mathcal{V}_3$ only has direct and bottom-up inference. For clarity of the method description, we discuss the general case with all three sources.

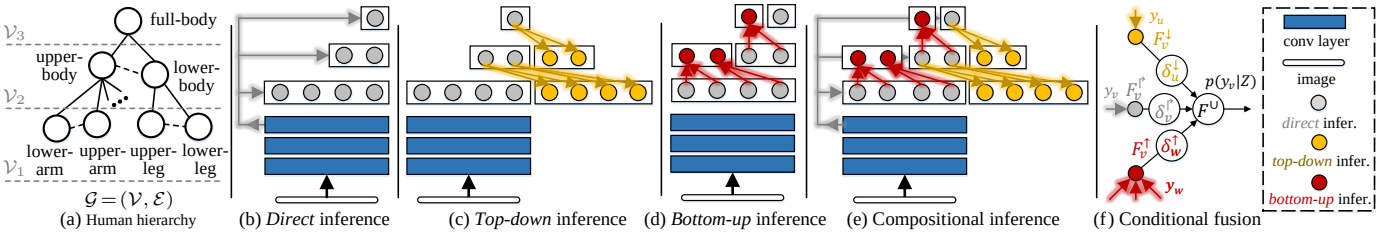


Fig. 3. Illustration of our conditional neural information fusion network for hierarchical human parsing. See §3.2.3 for details.

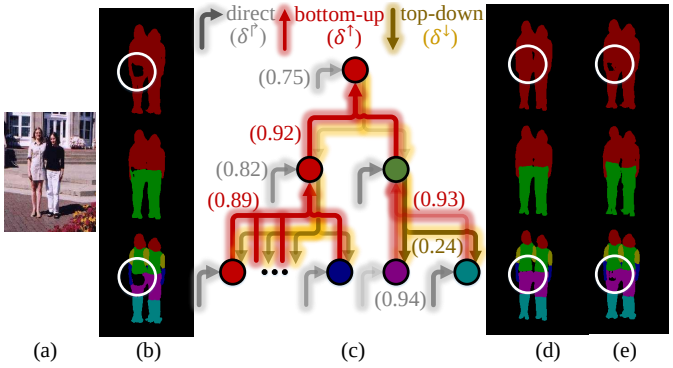


Fig. 4. Illustration of our compositional inference and conditional fusion (§3.2.3). (a) Input image. (b) Parsing results of direct inference. (c) Conditional information fusion, where the arrows with darker colors indicate higher values of gates δ. For clarity, in $(\cdot)$ we only show the gate values for a few inference processes. (d) Parsing results *w/* compositional inference and conditional fusion. (e) Parsing results of compositional inference only. The improved regions are highlighted in white circles.

where $\boldsymbol{C}_u^{\downarrow} \in \mathbb{R}^{1\times c}$ and $\boldsymbol{C}_{\boldsymbol{w}}^{\uparrow} \in \mathbb{R}^{1\times c|\boldsymbol{w}|}$. Specifically, for the bottom-up process, we concatenate all the child node embeddings $[\boldsymbol{h}_w]_{w\in\boldsymbol{w}} \in \mathbb{R}^{W\times H\times c|\boldsymbol{w}|}$. This means our decision is made upon the confidence of the union of the child nodes. Here, the confidence of a source can be viewed as a global score or statistic for interpreting the quality of the feature, which is learnt in an implicit manner.

Finally, for each node v, the fusion network $F_{\cup}$ combines the results from the three inference networks above for final prediction (see Fig. 3(e)):

$$\text{logit}(y_v|Z) = F^{\cup}(\delta_v^{\curvearrowright} F_v^{\curvearrowright}, \delta_u^{\downarrow} F_v^{\downarrow}, \delta_{\boldsymbol{w}}^{\uparrow} F_v^{\uparrow}; \boldsymbol{W}^{\cup}) \in \mathbb{R}_{\geq 0}^{W\times H}, \tag{17}$$

where $F^{\cup}: \mathbb{R}_{\geq 0}^{W\times H\times 3} \mapsto \mathbb{R}_{\geq 0}^{W\times H}$ is implemented by a small FCN, parameterized by $\boldsymbol{W}^{\cup}$. Fig. 4 provides an illustration of our conditional fusion process. As can be seen, δ provides a learnable gate mechanism that suggests how much information can be used from a source (see Fig. 4(c)). It is able to dynamically change the amount of information for different inference processes, *i.e.*, condition on the sources (see Fig. 4(d)). Thus, it yields better results than statically fusing the information with a weight-fixed fusion function (see Fig. 4(e)). More detailed studies of our conditional and compositional fusion can be found in §4.5.1.

Loss Function. To obtain the final l-level segmentation maps $\hat{\mathcal{Y}}_l = \{\hat{y}_v \in [0,1]^{W\times H}\}_{v\in\mathcal{V}_l}$, we apply *pixel-wise soft-max* (PSM) over the logits of l-level nodes: $\{\text{logit}(y_v|Z) \in \mathbb{R}_{\geq 0}^{W\times H}\}_{v\in\mathcal{V}_l}$. Thus, for each level, all the inference networks; $F^{\curvearrowright}, F^{\downarrow}, F^{\uparrow}$, and the fusion network $F^{\cup}$ are trained by the cross-entropy loss $\mathcal{L}_{\text{CE}}$ and the overall loss $\mathcal{L}_{\text{parsing}}$ is defined as:

$$\begin{aligned}\mathcal{L}_{\text{parsing}} = \sum\nolimits_{l=1}^{3} \big(&\mathcal{L}_{\text{CE}}(\hat{\mathcal{Y}}_l^{\curvearrowright}, \mathcal{Y}_l) + \mathcal{L}_{\text{CE}}(\hat{\mathcal{Y}}_l^{\downarrow}, \mathcal{Y}_l) + \\ &\mathcal{L}_{\text{CE}}(\hat{\mathcal{Y}}_l^{\uparrow}, \mathcal{Y}_l) + \mathcal{L}_{\text{CE}}(\hat{\mathcal{Y}}_l^{\cup}, \mathcal{Y}_l)\big).\end{aligned} \tag{18}$$

3.3 Part-Relation Aware Human Parser (PRHP)

Our CNIF framework provides a structured human parser that explores top-down and bottom-up context over human configurations. However, it has two limitations. **1)** CNIF only relies on compositional and decompositional relations which could be over generalized and simplified. It does not seem to characterize well the diverse part relations, especially the essential and distinct geometric constraints of different types of relations. **2)** Since information can diffuse in different directions, the human body yields a complex, cyclic graph topology. Hence an iterative inference is desirable for optimal result approximation. However, CNIF, as well as current arts, are built upon an immediate, *feed-forward* prediction scheme.

To address the above limitations, we further propose a part-relation aware human parser (PRHP), which explains part relations more comprehensively and precisely (§3.3.1) and efficiently reasons human semantics in an iterative, *feed-back* fashion (§3.3.2). With the human body hierarchy $\mathcal{G}$, PRHP is trained in a graph learning scheme, also using the full supervision from existing human parsing datasets.

3.3.1 Typed Human Part Relation Modeling

As mentioned in §1, there are diverse relations between human parts, *i.e.*, decomposition and composition between constituent and entire parts, and dependency between kinematically connected, same-level parts. One of the core ideas of PRHP is to model these relations in a type-wise manner, which provides an explicit bias towards capturing the specific geometric and anatomical constraints of different relations. For parts (nodes) u and v, their relation can be captured by the directed edge embedding $\boldsymbol{h}_{u,v}$, which is formulated as:

$$\boldsymbol{h}_{u,v} = R^r(F^r(\boldsymbol{h}_u), \boldsymbol{h}_v), \tag{19}$$

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 [41], 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.*,

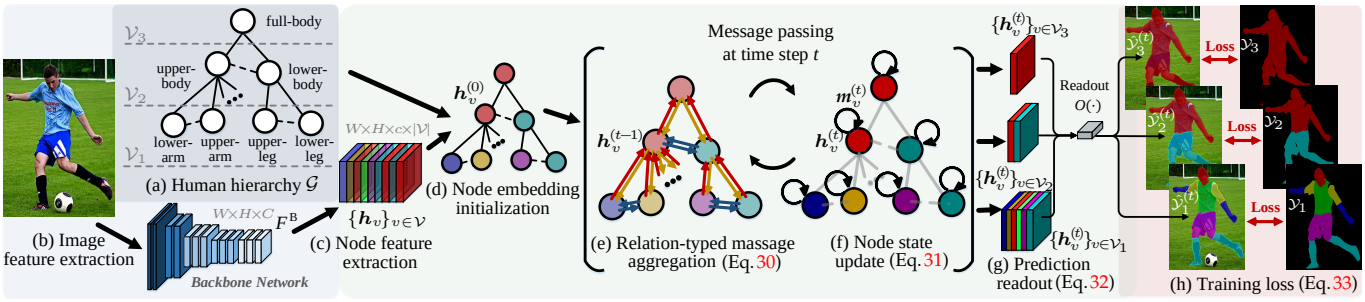


Fig. 5. Our PRHP model for hierarchical human parsing during training (§3.3). The main components in the flowchart are marked by (a)-(h).

$\boldsymbol{h}_{u,v}$) 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.

Decompositional Relation Modeling. Decompositional relations (full line:/ in Fig. 5(a)) are represented by those vertical edges starting from parent nodes to corresponding child nodes in the human body 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. 6):

$$\begin{aligned} \boldsymbol{h}_{u,v} &= R^{\text{dec}}(F^{\text{dec}}(\boldsymbol{h}_u), \boldsymbol{h}_v), \quad v \in \mathcal{C}_u, \\ F^{\text{dec}}(\boldsymbol{h}_u) &= \boldsymbol{h}_u \odot \texttt{att}^{\text{dec}}_{u,v}(\boldsymbol{h}_u). \end{aligned} \tag{20}$$

'$\odot$' 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:

$$\texttt{att}^{\text{dec}}_{u,v}(\boldsymbol{h}_u) = \text{PSM}([\phi^{\text{dec}}_v(\boldsymbol{h}_u)]_{v\in\mathcal{C}_u}) = \frac{\exp(\phi^{\text{dec}}_v(\boldsymbol{h}_u))}{\Sigma_{v'\in\mathcal{C}_u}\exp(\phi^{\text{dec}}_{v'}(\boldsymbol{h}_u))}, \tag{21}$$

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

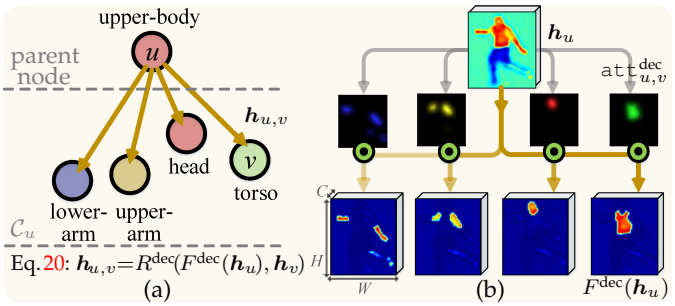


Fig. 6. Illustration of our decompositional relation modeling (§3.3.1). (a) Decompositional relations between the *upper-body* node (u) and its constituents ($\mathcal{C}_u$). (b) With the decompositional attentions $\{\texttt{att}^{\text{dec}}_{u,v}(\boldsymbol{h}_u)\}_{v\in\mathcal{C}_u}$, F^{dec} 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.

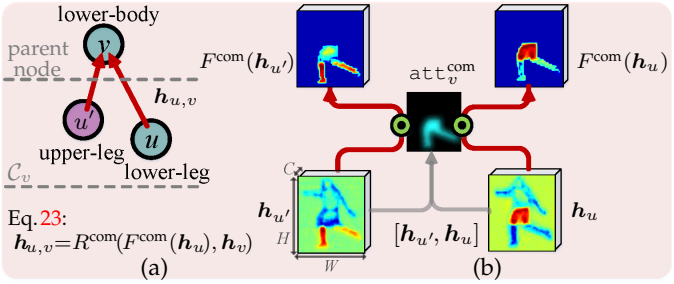


Fig. 7. Illustration of our compositional relation modeling (§3.3.1). (a) Compositional relations between the *lower-body* node (v) and its constituents ($\mathcal{C}_v$). (b) The compositional attention $\texttt{att}^{\text{com}}_v([\boldsymbol{h}_{u'}, \boldsymbol{h}_u])$ gathers information from all the constituents $\mathcal{C}_v$ and lets F^{com} enhance all the *lower-body* related features of $\mathcal{C}_v$.

$\sum_{v\in\mathcal{C}_u}\texttt{att}^{\text{dec}}_{u,v} = \mathbf{1}$, $\{\texttt{att}^{\text{dec}}_{u,v}(\boldsymbol{h}_u)\}_{v\in\mathcal{C}_u}$ 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. 6(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* $\{\texttt{att}^{\text{dec}}_{u,v}(\boldsymbol{h}_u)\}_{v\in\mathcal{C}_u} \in [0,1]^{W\times H\times|\mathcal{C}_u|}$:

$$\mathcal{L}_{\text{dec}} = \sum\nolimits_{u\in\mathcal{V}_2\cup\mathcal{V}_3} \mathcal{L}_{\text{CE}}\big(\{\texttt{att}^{\text{dec}}_{u,v}(\boldsymbol{h}_u)\}_{v\in\mathcal{C}_u}, \mathcal{Y}_{\mathcal{C}_u}\big). \tag{22}$$

Compositional Relation Modeling. In the human body 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. 7):

$$\begin{aligned} \boldsymbol{h}_{u,v} &= R^{\text{com}}(F^{\text{com}}(\boldsymbol{h}_u), \boldsymbol{h}_v), \quad u \in \mathcal{C}_v, \\ F^{\text{com}}(\boldsymbol{h}_u) &= \boldsymbol{h}_u \odot \texttt{att}^{\text{com}}_v([\boldsymbol{h}_{u'}]_{u'\in\mathcal{C}_v}). \end{aligned} \tag{23}$$

Here $\texttt{att}^{\text{com}}_v : \mathbb{R}^{W\times H\times c\times|\mathcal{C}_v|} \mapsto [0,1]^{W\times H}$ 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. 7(b)).

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:

$$\mathcal{L}_{\text{com}} = \sum\nolimits_{v\in\mathcal{V}_2\cup\mathcal{V}_3} \mathcal{L}_{\text{CE}}\big(\texttt{att}_v^{\text{com}}([\boldsymbol{h}_{u'}]_{u'\in\mathcal{C}_v}), y_v\big). \tag{24}$$

Dependency Relation Modeling. In $\mathcal{G}$, dependency relations are represented as horizontal edges (dashed line:-- in Fig. 5(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. 8):

$$\begin{aligned} \boldsymbol{h}_{u,v} &= R^{\text{dep}}(F^{\text{dep}}(\boldsymbol{h}_u), \boldsymbol{h}_v), \quad v \in \mathcal{K}_u, \\ F^{\text{dep}}(\boldsymbol{h}_u) &= F^{\text{cont}}(\boldsymbol{h}_u) \odot \texttt{att}_{u,v}^{\text{dep}}\big(F^{\text{cont}}(\boldsymbol{h}_u)\big), \end{aligned} \tag{25}$$

where $F^{\text{cont}}(\boldsymbol{h}_u) \in \mathbb{R}^{W\times H\times c}$ is used to extract the context of u, and $\texttt{att}_{u,v}^{\text{dep}}\big(F^{\text{cont}}(\boldsymbol{h}_u)\big) \in [0,1]^{W\times H}$ 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 [92], [93], the *context extraction* module F^{cont} is designed as:

$$\begin{aligned} F^{\text{cont}}(\boldsymbol{h}_u) &= \tau(\boldsymbol{h}_I \boldsymbol{A}^\top) \in \mathbb{R}^{W\times H\times c}, \\ \boldsymbol{A} &= \boldsymbol{h}_u'^\top \boldsymbol{W}^{\text{A}} \boldsymbol{h}_I' \in \mathbb{R}^{(WH)\times(WH)}, \end{aligned} \tag{26}$$

where $\boldsymbol{h}_u' \in \mathbb{R}^{(c+8)\times(WH)}$ and $\boldsymbol{h}_I' \in \mathbb{R}^{(C+8)\times(WH)}$ are node (part) and image representations augmented with spatial information, respectively, flattened into matrix formats. The last eight channels of $\boldsymbol{h}_u'$ and $\boldsymbol{h}_I'$ encode spatial coordinate information [94], 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 $1/H$. $\boldsymbol{W}^{\text{A}} \in \mathbb{R}^{(c+8)\times(C+8)}$ 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 $\boldsymbol{h}_I'$ for its context information. As a result, the affinity matrix $\boldsymbol{A}$ stores the attention 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 $\boldsymbol{h}_I$ with column-wise normalized weight matrix $\boldsymbol{A}^\top$: $\boldsymbol{h}_I \boldsymbol{A}^\top \in \mathbb{R}^{C\times(WH)}$. A 1×1 convolution based linear embedding function $\tau : \mathbb{R}^{W\times H\times C} \mapsto \mathbb{R}^{W\times H\times c}$ 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, $\texttt{att}_{u,v}^{\text{dep}}$ is defined as:

$$\texttt{att}_{u,v}^{\text{dep}}\big(F^{\text{cont}}(\boldsymbol{h}_u)\big) = \text{PSM}\big([\phi_v^{\text{dep}}(\boldsymbol{h}_u)]_{v\in\mathcal{K}_u}\big). \tag{27}$$

Here $\phi_v^{\text{dep}}(\cdot) \in \mathbb{R}^{W\times H}$ gives an importance map for v, using a 1×1 convolutional layer. Through the *pixel-wise soft-max* operation $\text{PSM}(\cdot)$, we enforce $\sum_{v\in\mathcal{K}_u} \texttt{att}_{u,v}^{\text{dep}} = \mathbf{1}$, 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:

$$\mathcal{L}_{\text{dep}} = \sum\nolimits_{u\in\mathcal{V}_1\cup\mathcal{V}_2} \mathcal{L}_{\text{CE}}\big(\{\texttt{att}_{u,v}^{\text{dep}}(\boldsymbol{h}_u)\}_{v\in\mathcal{K}_u}, \mathcal{Y}_{\mathcal{K}_u}\big), \tag{28}$$

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.

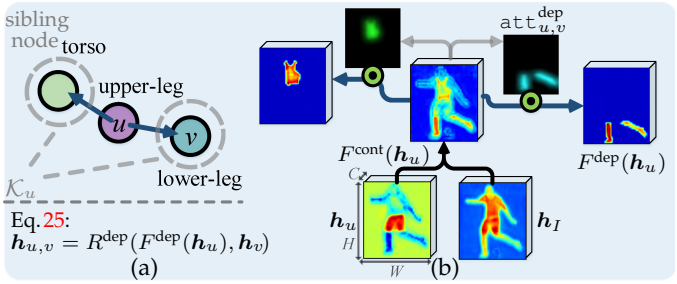


Fig. 8. Illustration of our dependency relation modeling (§3.3.1). (a) Dependency relations between the *upper-body* node (u) and its siblings ($\mathcal{K}_u$). (b) The dependency attention $\{\texttt{att}_{u,v}^{\text{dep}}(F^{\text{cont}}(\boldsymbol{h}_u))\}_{v\in\mathcal{K}_u}$, derived from u's contextual information $F^{\text{cont}}(\boldsymbol{h}_u)$, gives separate importance for different siblings $\mathcal{K}_u$.

3.3.2 Iterative Inference over the Human Body Hierarchy

Human bodies present a hierarchical structure that yields directed and cyclic inference graphs. For such problems, approximate iterative inference algorithms are often adopted [40], [41]. Compared with the feed-forward network architectures adapted in prior arts, iterative algorithms could offer a more favorable solution in such a structured setting, *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 [41]. 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 $\boldsymbol{h}_v$:

$$\begin{aligned} \boldsymbol{m}_v^{(t)} &= \sum\nolimits_{u\in\mathcal{N}_v} M(\boldsymbol{h}_u^{(t-1)}, \boldsymbol{h}_v^{(t-1)}), \\ \boldsymbol{h}_v^{(t)} &= U(\boldsymbol{h}_v^{(t-1)}, \boldsymbol{m}_v^{(t)}), \end{aligned} \tag{29}$$

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. 5(e)-(f)):

$$\boldsymbol{m}_v^{(t)} = \underbrace{\sum\nolimits_{u\in\mathcal{P}_v} \boldsymbol{h}_{u,v}^{(t-1)}}_{\text{decomposition}} + \underbrace{\sum\nolimits_{u\in\mathcal{C}_v} \boldsymbol{h}_{u,v}^{(t-1)}}_{\text{composition}} + \underbrace{\sum\nolimits_{u\in\mathcal{K}_v} \boldsymbol{h}_{u,v}^{(t-1)}}_{\text{dependency}}, \tag{30}$$

$$\boldsymbol{h}_v^{(t)} = U_{\text{convGRU}}(\boldsymbol{h}_v^{(t-1)}, \boldsymbol{m}_v^{(t)}), \tag{31}$$

where the initial state $\boldsymbol{h}_v^{(0)}$ is obtained by Eq. 10. Here, the message aggregation step (Eq. 30) 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. 31, we use a convGRU [95], 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.

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. 5(g)), and *pixel-wise soft-max* (PSM) for normalization:

$$\hat{\mathcal{Y}}_l^{(t)} = \{\hat{y}_v^{(t)}\}_{v\in\mathcal{V}_l} = \text{PSM}\big([O(\boldsymbol{h}_v^{(t)})]_{v\in\mathcal{V}_l}\big). \tag{32}$$

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. 5(h)):

$$\mathcal{L}_{\text{parsing}}^{(t)} = \sum\nolimits_{l=1}^{3} \mathcal{L}_{\text{CE}}^{(t)}(\hat{\mathcal{Y}}_l^{(t)}, \mathcal{Y}_l). \tag{33}$$

Considering Eqs. 22, 24, 28, and 33, the overall loss is defined as:

$$\mathcal{L} = \sum\nolimits_{t=1}^{T} \big(\mathcal{L}_{\text{parsing}}^{(t)} + \alpha(\mathcal{L}_{\text{com}}^{(t)} + \mathcal{L}_{\text{dec}}^{(t)} + \mathcal{L}_{\text{dep}}^{(t)})\big), \tag{34}$$

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 §4.5.2.

In the supplementary material, we provide pseudo-code descriptions for our proposed two parsers.

3.4 Implementation Details

3.4.1 CNIF based Human Parser

Backbone Network. Our feature extraction network F^{B} in Eq. 8 uses the convolutional blocks of ResNet101 [96]. The stride is set to 16, *i.e.*, the resolution of the output is 1/16 of that of the input, for high computational efficiency. In addition, the ASPP module [97] is applied for extracting more powerful features with multi-scale context. The ASPP-enhanced feature is compressed by a 1×1 convolutional layer with *ReLU* activation. The compressed 512-*d* feature is further $\times2$ upsampled and element-wisely added with the feature from the second convolutional block of ResNet101, to encode more spatial details. Thus, given an input image I the feature extraction network F^{B} produces a 512-*d* image representation $\boldsymbol{h}_I$ whose spatial dimensions are 1/8 of I.

Node Embedding. We implement F_l^{LSF} (Eq. 9) using a 3×3 convolutional layer with Batch Normalization (BN) and *ReLU* activation, whose parameters are shared by all the nodes located in the l-th level. This is used for extracting specific features $\{\boldsymbol{h}_1^{\text{LSF}}, \boldsymbol{h}_2^{\text{LSF}}, \boldsymbol{h}_3^{\text{LSF}}\}$ for the three semantic-levels. For each node v, an independent SE [90] block, F_v^{SE} in Eq. 10, is further applied to extract its specific embedding $\boldsymbol{h}_v$ with an extremely light-weight architecture. In addition, we set the channel size of level-specific embeddings and node features $c=64$ to maintain high computational efficiency.

Direct Inference Network. In Eq. 11, the direct inference network takes node embeddings as inputs and predicts segmentation maps. $F^{\uparrow}$ is implemented by a stack of three 1×1 convolutional layers.

Top-down/Bottom-up Inference Network. The architectures of the top-down $F^{\downarrow}$ (Eq. 12) and bottom-up $F^{\uparrow}$ (Eq. 14) inference networks are very similar, and only differ in their strategies of processing the input features (see Eq. 13). Both are achieved by three cascaded convolutional layers, with convolution sizes of 3×3, 3×3 and 1×1, respectively.

Information Fusion Network. $F^{\cup}$ in Eq. 17 consists of three 1×1 convolutional layers with *ReLU* activations for non-linear mapping, where the first two aim to aggregate the information from different sources, while the final one is to generate the final prediction.

3.4.2 PRHP Model

Relation Networks. Each typed relation network R^r in Eq. 19 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. 31, the update function U_{convGRU} is implemented by a convolutional GRU with 3×3 convolution kernels. The readout function O in Eq. 32 applies a 1×1 convolution operation over the node embeddings. 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.

All the units of our PRHP model are built on convolution operations, leading to spatial information preservation.

4 Experiments

4.1 Experimental Settings

Datasets. We perform experiments on the following datasets:

- **LIP [6]** has 50,462 single-person images with elaborate pixel-wise annotations of 19 part categories (*e.g.*, *hair*, *face*, *left-/right-arms*, *left-/right-legs*, *left-/right-shoes*, *etc.*). The images are divided into 30,462 samples for training, 10,000 for validation and 10,000 for testing.
- **PASCAL-Person-Part [7]** contains 3,533 multi-person images with challenging poses and viewpoints (1,716 for training and 1,817 for testing). It provides careful pixel-wise annotations for six body parts (*i.e.*, *head, torso, upper-/lower-arms*, and *upper-/lower-legs*).
- **ATR [18]** includes 7,700 single-person images (6,000 for training, 700 for validation and 1,000 for testing), annotated at pixel-level with 17 categories, *e.g.*, *hat, sunglass, face, upper-clothes, pants, left-/right-arms, left-/right-legs, etc.*
- **Fashion Clothing [19]** consists of Colorful Fashion Parsing [53], Fashionista [54], and Clothing Co-Parsing [55]. It is more concerned with human clothing details, including 17 categories (*e.g.*, *glass, hair, pants, shoes, shirt, upper-clothes, skirt, scarf, socks, etc.*). It has 4,371 images in total (3,934 for training, and 437 for testing).
- **PPSS [68]** has 3,673 samples, collected from 171 surveillance videos containing diverse general challenges (*i.e.*, occlusion, illumination variation) in real-word scenes. PPSS is divided into 1,781 and 1,892 images for training and testing, respectively. Pixel-wise annotations for *hair, face, upper-/lower-clothes, arm*, and *leg* are provided.

Evaluation Metrics. For LIP, following its standard protocol [71], we report pixel accuracy, mean accuracy and mean Intersection-over-Union (mIoU). For PASCAL-Person-Part, following conventions [7], [70], [74], the performance is evaluated in terms of mIoU. For ATR and Fashion Clothing, we report five metrics as [19] does, including pixel accuracy, foreground accuracy, average precision, recall, and F1-score.

Training. Our method is implemented on PyTorch and trained on four NVIDIA Tesla V100 GPUs with a 32GB memory per-card. During training, the weights of the backbone network are loaded from ImageNet pre-trained

TABLE 1
Comparison results on LIP `val` [6]. † indicates extra pose information used. (The two best scores are marked in **red** and **blue**, respectively. These notes are the same for Tables 2,3,4, and 5.)

Method	pixAcc.	Mean Acc.	Mean IoU
SegNet [100]	69.04	24.00	18.17
FCN-8s [3]	76.06	36.75	28.29
DeepLabV2 [4]	82.66	51.64	41.64
Attention [69]	83.43	54.39	42.92
†Attention+SSL [6]	84.36	54.94	44.73
DeepLabV3+ [101]	84.09	55.62	44.80
ASN [102]	-	-	45.41
†SSL [6]	-	-	46.19
MMAN [74]	85.24	57.60	46.93
†SS-NAN [71]	87.59	56.03	47.92
HSP-PRI [103]	85.07	60.54	48.16
†MuLA [8]	**88.5**	60.5	49.3
PSPNet [99]	86.23	61.33	50.56
CE2P [82]	87.37	63.20	53.10
BraidNet [104]	87.60	66.09	54.42
CNIF (**Ours**)	88.03	**68.80**	**57.74**
PRHP (**Ours**)	**89.05**	**70.58**	**59.25**

TABLE 2
Per-class comparison results on PASCAL-Person-Part `test` [7].

Method	Head	Torso	U-Arm	L-Arm	U-Leg	L-Leg	B.G.	Ave.
HAZN [70]	80.79	59.11	43.05	42.76	38.99	34.46	93.59	56.11
Attention [69]	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 [6]	83.26	62.40	47.80	45.58	42.32	39.48	94.68	59.36
Attention+MMAN [74]	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 [71]	86.43	67.28	51.09	48.07	44.82	42.15	97.23	62.44
Structure LSTM [105]	82.89	67.15	51.42	48.72	51.72	45.91	97.18	63.57
Joint [7]	85.50	67.87	54.72	54.30	48.25	44.76	95.32	64.39
DeepLabV2 [4]	-	-	-	-	-	-	-	64.94
†MuLA [8]	84.6	68.3	57.5	54.1	49.6	46.4	95.6	65.1
PCNet [34]	86.81	69.06	55.35	55.27	50.21	48.54	96.07	65.90
Holistic [80]	86.00	69.85	56.63	55.92	51.46	48.82	95.73	66.34
WSHP [5]	87.15	72.28	57.07	56.21	52.43	50.36	**97.72**	67.60
DeepLabV3+ [101]	87.02	72.02	60.37	57.36	53.54	48.52	96.07	67.84
SPGNet [106]	87.67	71.41	61.69	60.35	52.62	48.80	95.98	68.36
PGN [98]	**90.89**	**75.12**	55.83	**64.61**	55.42	41.57	95.33	68.40
CNIF (**Ours**)	88.02	72.91	**64.31**	63.52	**55.61**	**54.96**	96.02	**70.76**
PRHP (**Ours**)	**89.73**	**75.22**	**66.87**	**66.21**	**58.69**	**58.17**	**96.94**	**73.12**

TABLE 3
Comparison results on ATR `test` [18].

Method	pixAcc.	F.G. Acc.	Prec.	Recall	F-1
Yamaguchi [54]	84.38	55.59	37.54	51.05	41.80
Paperdoll [61]	88.96	62.18	52.75	49.43	44.76
M-CNN [66]	89.57	73.98	64.56	65.17	62.81
ATR [18]	91.11	71.04	71.69	60.25	64.38
DeepLabV2 [4]	94.42	82.93	78.48	69.24	73.53
PSPNet [99]	95.20	80.23	79.66	73.79	75.84
Attention [69]	95.41	85.71	81.30	73.55	77.23
DeepLabV3+ [101]	95.96	83.04	80.41	78.79	79.49
Co-CNN [67]	96.02	83.57	**84.95**	77.66	80.14
LG-LSTM [33]	96.18	84.79	84.64	79.43	80.97
TGPNet [19]	**96.45**	**87.91**	83.36	80.22	81.76
CNIF (**Ours**)	96.26	**87.91**	84.62	**86.41**	**85.51**
PRHP (**Ours**)	**96.84**	**89.23**	**86.17**	**88.35**	**87.25**

ResNet101 [96]. We train our model on the five aforementioned datasets with their respective training samples, separately. For data preparation, following [82], [98], we apply data augmentation techniques for all the training samples, *e.g.*, scaling, cropping and left-right flipping. The random scale is set from 0.5 to 2.0, while the crop size is set to 473×473. For optimization, we adopt SGD with a momentum of 0.9, and weight_decay of 0.0005. For the learning rate, we use the 'poly' learning rate schedule [4], [99], $lr = base\_lr \times (1 - \frac{iters}{total\_iters})^{power}$, in which $power$=0.9 and $base\_lr$=0.007. The $total\_iters$ is $epochs \times batch\_size$, where $batch\_size$=40 and $epochs$=150.

Testing. All the testing procedures are carried out on a single NVIDIA TITAN Xp GPU with 12GB memory. For each test sample, we set the long side of the image to 473 pixels and maintain the original aspect ratio. Following conventions [8], [99], we average the per-pixel classification scores at multi-scales with flipping, *i.e.*, the scale is 0.5 to 1.5 (in increments of 0.25) times the original size.

4.2 Quantitative Results

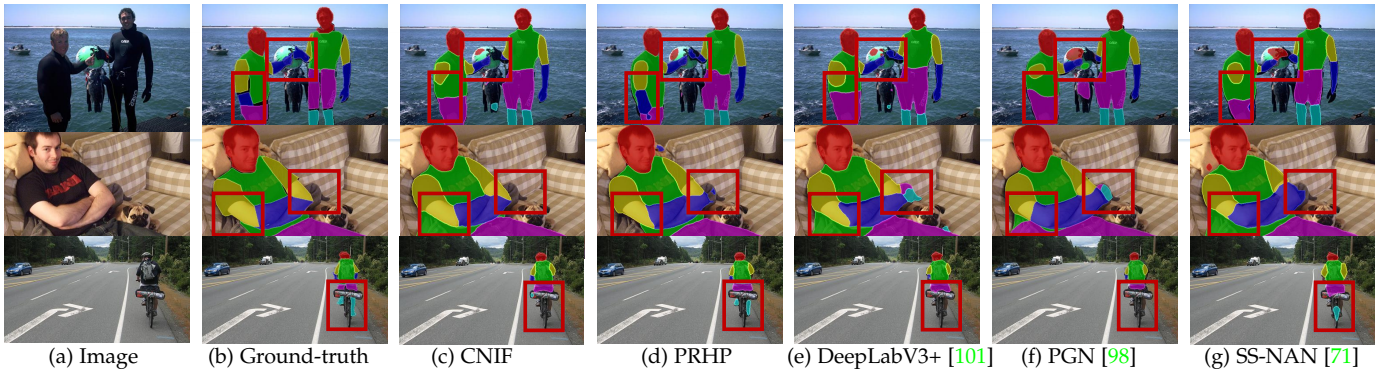
We compare the proposed two parsers with several strong baselines on the five aforementioned challenging datasets. Note that a recent method [76] is not included, as it is simultaneously trained on three datasets (*i.e.*, PASCAL-Person-Part [7], ATR [18], and CIHP [98]) with transfer learning.

LIP [6]. LIP is a gold standard benchmark for human parsing. We compare our method with 15 state-of-the-arts on LIP `val` set in Table 1. We first find that general semantic segmentation methods [3], [4], [100], [101] tend to perform worse than specifically designed human parsers. This indicates the importance of reasoning human structures in this problem. In addition, though recent human parsers gain impressive results, our two models still outperform all the competitors by a large margin. For instance, our CNIF based parser achieves a huge boost in average IoU (3.32% better than the second best method, BraidNet [104] and 4.64% better than the third best, CE2P [82]). In terms of pixAcc., mean Acc., and mean IoU, our PRHP model further surpasses CNIF by 1.02%, 1.78% and 1.51%, respectively. We would also like to mention that our parsers do not use additional pose [5]–[8] or edge [82] information.

PASCAL-Person-Part [7]. In Table 2, we compare our models against 17 recent methods on PASCAL-Person-Part `test`. From the results, we can again see that our approaches outperform previous methods across the vast majority of classes and on average. Specifically, our CNIF and PRHP outperforms the prior best, PGN [98], by 2.36% and 4.72%, respectively, in terms of *mIoU*. Such performance gains are particularly impressive considering that improvement on this dataset is very challenging.

ATR [18]. Table 3 presents comparisons with 11 previous methods on ATR `test` set. Note that [105] is not included in comparison, because it makes use of extra 10,000 images from [67] for training. Our approaches set new state-of-the-arts for all five metrics, outperforming all other methods by a large margin. For example, CNIF achieves an average F1-score of 85.51%, which is 3.75% better than TGPNet [19] and 4.54% better than LG-LSTM [33]. And our PRHP further provides a more considerable performance gain in F1-score, *i.e.*, 5.49% and 6.28% higher than TGPNet [19] and LG-LSTM [33], respectively.

Fashion Clothing [19]. The quantitative comparison results with five competitors on Fashion Clothing `test` are summarized in Table 4, where we take the pre-computed evaluation from [19]. Our models surpass other competitors across all metrics by large margins. Notably, our CNIF and PRHP yield F1-scores of 58.12% and 60.19%, respectively, while



(a) Image (b) Ground-truth (c) CNIF (d) PRHP (e) DeepLabV3+ [101] (f) PGN [98] (g) SS-NAN [71]

Fig. 9. Visual comparison on PASCAL-Person-Part `test` [7]. Our CNIF (c) and PRHP (d) generate more accurate predictions, compared to other famous methods [71], [98], [101] (e-g). The improved labeled results by our parser are denoted in red boxes. See §4.3 for details.

TABLE 4
Comparison results on Fashion Clothing `test` [19].

Method	pixAcc.	F.G. Acc.	Prec.	Recall	F-1
Yamaguchi [54]	81.32	32.24	23.74	23.68	22.67
Paperdoll [61]	87.17	50.59	45.80	34.20	35.13
DeepLabV2 [4]	87.68	56.08	35.35	39.00	37.09
Attention [69]	90.58	64.47	47.11	50.35	48.68
TGPNet [19]	91.25	66.37	50.71	53.18	51.92
CNIF (**Ours**)	**92.20**	**68.59**	**56.84**	**59.47**	**58.12**
PRHP (**Ours**)	**93.12**	**70.57**	**58.73**	**61.72**	**60.19**

TABLE 5
Comparison results on PPSS `test` [68].

Method	Head	Face	U-Cloth	Arms	L-Cloth	Legs	B.G.	Ave.
DL [68]	22.0	29.1	57.3	10.6	46.1	12.9	68.6	35.2
DDN [68]	35.5	44.1	68.4	17.0	61.7	23.8	80.0	47.2
ASN [102]	51.7	51.0	65.9	29.5	52.8	20.3	83.8	50.7
MMAN [74]	53.1	50.2	69.0	29.4	55.9	21.4	85.7	52.1
LCPC [107]	55.6	46.6	71.9	30.9	58.8	24.6	86.2	53.5
CNIF (**Ours**)	**67.6**	**60.8**	**80.8**	**46.8**	**69.5**	**28.7**	**90.6**	**60.5**
PRHP (**Ours**)	**68.8**	**63.2**	**81.7**	**49.3**	**70.8**	**32.0**	**91.4**	**65.3**

those for TGPNet [19] is 51.92%.

PPSS [68]. Table 5 compares our methods against five famous methods on PPSS `test` set. The evaluation results demonstrate that our CNIF yields an mIoU of 60.5%, while those of MMAN [74] and LCPC [107] are 52.1% and 53.5%, respectively. In addition, our PRHP achieves 65.3% mIoU, with substantial gains over MMAN [74] and LCPC [107] of 13.2% and 11.8%, respectively.

4.3 Qualitative Results

Some qualitative comparison results on PASCAL-Person-Part `test` are depicted in Fig. 9. We can see that our approaches output more precise parsing results than other competitors [71], [98], [101], 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 PRHP parser still correctly recognizes the confusing parts of the person in the middle.

4.4 Runtime Comparison

Either CNIF or PRHP does not require any other pre-/post-processing steps (*i.e.*, over-segmentation [32], [105], human pose [7], CRF [7]). Thus CNIF and PRHP achieve high processing speed of 23.0fps and 12.0fps, respectively (averaged on PASCAL-Person-Part). This is faster than or on par with prior deep human parsers, such as Joint [7] (0.1fps), Attention+SSL [6] (2.0fps), MMAN [74] (3.5fps), MuLA [8] (15fps), SS-NAN [71] (2.0fps), and LG-LSTM [33] (3.0fps).

TABLE 6
Ablation study (§4.5) for CNIF model on PASCAL-Person-Part `test` [7].

Aspects	Module	mIoU $\mathcal{V}_1$	mIoU $\mathcal{V}_2$	mIoU $\mathcal{V}_3$
CNIF	direct + bottom-up + top-down + conditional fusion	70.76	81.62	91.31
Backbone	direct infer. *w/o* hierarchy	64.14	-	-
Variant	direct	65.27	77.83	88.29
	direct + bottom-up	65.42	78.37	90.10
	direct + top-down	69.02	78.91	88.40
	direct + bottom-up + top-down	69.43	80.34	91.02

4.5 Diagnostic Experiment

To demonstrate how each component in our parsers contributes to the performance, a series of ablation experiments are conducted on PASCAL-Person-Part `test` [7] using mIoU metric. As PRHP provides a more general and powerful form of CNIF based parser, we first quantity the effectiveness of each essential ingredient of CNIF (§4.5.1), and then provide in-depth analyses of PRHP (§4.5.2). This would better verify our main points and coincide with the experiments in prior conference papers. The training and evaluation followed the same protocol as in §4.1.

4.5.1 Ablation Study for CNIF based Human Parser

Table 6 shows the evaluation of our CNIF model compared to ablated versions without certain key components. Here, $\mathcal{V}_1$ denotes the automatic parts (*e.g.*, *head, leg, etc.*), $\mathcal{V}_3$ *lower-/upper body*, and $\mathcal{V}_3$ *full body*. All the variants are retrained independently with their specific network architectures.

Hierarchical Human Semantic Parsing. Instead of only modeling the fine-grained parts in $\mathcal{V}^1$ (*i.e.*, *backbone*), even directly learning to parse the whole human body hierarchy (*i.e.*, *direct*) can bring a performance gain (64.14→65.27). This suggests that modeling the human body hierarchy leads to a comprehensive understanding of human semantics.

Fusing Information from Different Inference Processes. Considering the last four rows in Table 6, we can find that

Fig. 10. Visualizations of typical failure cases on PASCAL-Person-Part `test` set [7].

TABLE 7
Ablation study (§4.5) for PRHP model on PASCAL-Person-Part `test` [7].

Aspects	Module	mIoU ($\mathcal{V}_1$)
PRHP	**full model** (composition+decomposition+ dependency + 2 iterations)	73.12
Relation modeling	type-agnostic	70.37
	type-specific *w/o* F^r	71.65
	decomposition relation	71.38
	composition relation	69.35
	dependency relation	69.43
Iterative Inference T	0 iteration	68.84
	1 iteration	72.17
	3 iterations	73.19
	4 iterations	73.22
	5 iterations	73.23

further integrating bottom-up and top-down inference provides substantial performance gain. This demonstrates the benefit of exploiting human structures and efficient information fusion strategies in this problem. Note that in (*direct* vs. *direct+bottom-up*) and (*direct+top-down* vs. *direct+bottom-up+top-down*), even for the 1st-level nodes that do not have bottom-up inference, the training itself brings performance gain. The reason is that the bottom-up inference explicitly captures compositional relations and thus improves the quality of the learnt features. Similar observations can also be found in (*direct* vs. *direct+top-down*) and (*direct+bottom-up* vs. *direct+bottom-up+top-down*) for the 3rd-level node. These observations suggest the compositional information fusion not only improves the predictions during inference but also boosts the learning ability of our human parser model.

Conditional Information Fusion. Comparing the performance of our full CNIF model and *direct+bottom-up+top-down* baseline, we can conclude that conditionally fusing information boosts performance, as the information from low-quality sources can be suppressed. This also provides a new glimpse into the information fusion mechanism over hierarchical models.

4.5.2 Ablation Study for PRHP Model

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 five variant models, as listed in Table 6: **1)** 'Type-agnostic' shows the performance when modeling different human part relations in a type-agnostic manner: $\boldsymbol{h}_{u,v}{=}R([\boldsymbol{h}_u, \boldsymbol{h}_v])$; **2)** 'Type-specific *w/o* F^r' gives the performance without the relation-adaption operation F^r in Eq. 19: $\boldsymbol{h}_{u,v}{=}R^r([\boldsymbol{h}_u, \boldsymbol{h}_v])$; **3-5)** '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. 19). Three main conclusions can be drawn: **1)** Typed relation modeling leads to more effective human structure learning, as 'Type-specific *w/o* F^r' improves 'Type-agnostic' by 1.28%. **2)** Exploring different kinds of relations are meaningful, as our full model considering all the three kinds of relations achieves the best performance. **3)** 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. 30 and Eq. 31. 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 $\mathcal{V}_3$, 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.

4.6 Failure Case Analysis

To give a deeper insight into our methods, we present two representative failure cases in Fig. 10. As seen, our proposed model faces difficulties with low-quality images or dim scenes. Besides, our method may produce inferior results for humans at very small scales. In the future, we will therefore focus on addressing these issues.

5 Conclusion

In the human semantic parsing task, structure modeling is an essential, albeit inherently difficult, avenue to explore. In this work, we parse human parts in a hierarchical form, enabling us to capture human semantics from a more comprehensive view. We first tackle this problem through a neural information fusion framework that explores and combines the information from the direct, top-down, and bottom-up inference processes, while considering the reliability of each process. Based on this, we further address relation modeling/reasoning 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, we make convolutional, message passing based approximations, which enjoys the advantages of iterative optimization and spatial information preservation. Extensive quantitative and qualitative comparisons performed on five datasets demonstrate that our methods outperform all other competitors.

ACKNOWLEDGMENT

This work was supported partially by CCF-Baidu Open Fund and Zhejiang Lab's Open Fund (No. 2019KD0AB04).

REFERENCES

[1] W. Wang, Z. Zhang, S. Qi, J. Shen, Y. Pang, and L. Shao, "Learning compositional neural information fusion for human parsing," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2019, pp. 5703–5713.

[2] W. Wang, H. Zhu, J. Dai, Y. Pang, J. Shen, and L. Shao, "Hierarchical human parsing with typed part-relation reasoning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 8929–8939.

[3] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 3431–3440.

[4] 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 Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 4, pp. 834–848, 2018.

[5] 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 *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 70–78.

[6] 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 *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 932–940.

[7] F. Xia, P. Wang, X. Chen, and A. L. Yuille, "Joint multi-person pose estimation and semantic part segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 6769–6778.

[8] X. Nie, J. Feng, and S. Yan, "Mutual learning to adapt for joint human parsing and pose estimation," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 502–517.

[9] R. Kimchi, "Primacy of wholistic processing and global/local paradigm: a critical review." *Psychological Bulletin*, p. 24, 1992.

[10] D. Navon, "Forest before trees: The precedence of global features in visual perception," *Cognitive Psychology*, pp. 353–383, 1977.

[11] A. J. Marcel, "Conscious and unconscious perception: An approach to the relations between phenomenal experience and perceptual processes," *Cognitive psychology*, vol. 15, no. 2, pp. 238–300, 1983.

[12] T. Wu and S.-C. Zhu, "A numerical study of the bottom-up and top-down inference processes in and-or graphs," *Int. J. Comput. Vis.*, vol. 93, no. 2, pp. 226–252, 2011.

[13] K. Grill-Spector and K. S. Weiner, "The functional architecture of the ventral temporal cortex and its role in categorization," *Nature Reviews Neuroscience*, vol. 15, no. 8, p. 536, 2014.

[14] W. Wang, J. Shen, M.-M. Cheng, and L. Shao, "An iterative and cooperative top-down and bottom-up inference network for salient object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019.

[15] B. Epshtein, I. Lifshitz, and S. Ullman, "Image interpretation by a single bottom-up top-down cycle," *Proceedings of the National Academy of Sciences*, vol. 105, no. 38, pp. 14 298–14 303, 2008.

[16] L. Zhu, Y. Chen, Y. Lu, C. Lin, and A. Yuille, "Max margin and/or graph learning for parsing the human body," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2008, pp. 1–8.

[17] W. Wang, Y. Xu, J. Shen, and S.-C. Zhu, "Attentive fashion grammar network for fashion landmark detection and clothing category classification," in *CVPR*, 2018, pp. 4271–4280.

[18] 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 Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 12, pp. 2402–2414, 2015.

[19] X. Luo, Z. Su, J. Guo, G. Zhang, and X. He, "Trusted guidance pyramid network for human parsing," in *ACM MM*, 2018, pp. 654–662.

[20] Z. S. Harris, "Distributional structure," *Word*, vol. 10, no. 2-3, pp. 146–162, 1954.

[21] R. Mottaghi, X. Chen, X. Liu, N.-G. Cho, S.-W. Lee, S. Fidler, R. Urtasun, and A. Yuille, "The role of context for object detection and semantic segmentation in the wild," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2014, pp. 891–898.

[22] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan, "Object detection with discriminatively trained part-based models," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 9, pp. 1627–1645, 2010.

[23] P. F. Felzenszwalb and D. P. Huttenlocher, "Pictorial structures for object recognition," *Int. J. Comput. Vis.*, vol. 61, no. 1, pp. 55–79, 2005.

[24] T. Joachims, T. Finley, and C.-N. J. Yu, "Cutting-plane training of structural svms," *Machine Learning*, vol. 77, no. 1, pp. 27–59, 2009.

[25] X. Song, T. Wu, Y. Jia, and S.-C. Zhu, "Discriminatively trained and-or tree models for object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2013, pp. 3278–3285.

[26] A. Jain, A. R. Zamir, S. Savarese, and A. Saxena, "Structural-RNN: Deep learning on spatio-temporal graphs," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 5308–5317.

[27] H.-S. Fang, Y. Xu, W. Wang, X. Liu, and S.-C. Zhu, "Learning pose grammar to encode human body configuration for 3d pose estimation," in *The AAAI Conference on Artificial Intelligence*, 2018, pp. 6821–6828.

[28] S. Behnke, *Hierarchical neural networks for image interpretation*. Springer, 2003, vol. 2766.

[29] J. Gilmer, S. S. Schoenholz, P. F. Riley, O. Vinyals, and G. E. Dahl, "Neural message passing for quantum chemistry," in *International Conference on Machine Learning*, 2017, pp. 1263–1272.

[30] S. Zheng, S. Jayasumana, B. Romera-Paredes, V. Vineet, Z. Su, D. Du, C. Huang, and P. H. Torr, "Conditional random fields as recurrent neural networks," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2015, pp. 1529–1537.

[31] S. Qi, W. Wang, B. Jia, J. Shen, and S.-C. Zhu, "Learning human-object interactions by graph parsing neural networks," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 401–417.

[32] X. Liang, X. Shen, J. Feng, L. Lin, and S. Yan, "Semantic object parsing with graph lstm," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 125–143.

[33] X. Liang, X. Shen, D. Xiang, J. Feng, L. Lin, and S. Yan, "Semantic object parsing with local-global long short-term memory," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 3185–3193.

[34] B. Zhu, Y. Chen, M. Tang, and J. Wang, "Progressive cognitive human parsing," in *The AAAI Conference on Artificial Intelligence*, 2018, pp. 7607–7614.

[35] W. Tang, P. Yu, and Y. Wu, "Deeply learned compositional models for human pose estimation," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 190–206.

[36] Z. Zheng, W. Wang, S. Qi, and S.-C. Zhu, "Reasoning visual dialogs with structural and partial observations," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 6669–6678.

[37] B. Khaleghi, A. Khamis, F. O. Karray, and S. N. Razavi, "Multisensor data fusion: A review of the state-of-the-art," *Information Fusion*, vol. 14, no. 1, pp. 28–44, 2013.

[38] G. E. Hinton, "Products of experts," in *International Conference on Artificial Neural Networks*, 1999.

[39] T. G. Dietterich, "Ensemble methods in machine learning," in *International workshop on multiple classifier systems*, 2000, pp. 1–15.

[40] M. J. Wainwright, M. I. Jordan *et al.*, "Graphical models, exponential families, and variational inference," *Foundations and Trends® in Machine Learning*, pp. 1–305, 2008.

[41] F. Scarselli, M. Gori, A. C. Tsoi, M. Hagenbuchner, and G. Monfardini, "The graph neural network model," *IEEE Transactions on Neural Networks*, vol. 20, no. 1, pp. 61–80, 2008.

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

[43] 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 *Proc. Advances Neural Inf. Process. Syst.*, 2015, pp. 2224–2232.

[44] M. Niepert, M. Ahmed, and K. Kutzkov, "Learning convolutional neural networks for graphs," in *International Conference on Machine Learning*, 2016, pp. 2014–2023.

[45] T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," in *Proc. Int. Conf. Learn. Representations*, 2017.

[46] P. Velickovic, G. Cucurull, A. Casanova, A. Romero, P. Lio, and Y. Bengio, "Graph attention networks," in *Proc. Int. Conf. Learn. Representations*, 2018.

[47] W. Wang, X. Lu, J. Shen, D. J. Crandall, and L. Shao, "Zero-shot video object segmentation via attentive graph neural networks," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2019, pp. 9236–9245.

[48] L. Fan, Y. Chen, P. Wei, W. Wang, and S.-C. Zhu, "Inferring shared attention in social scene videos," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 6460–6468.

[49] L. Fan, W. Wang, S. Huang, X. Tang, and S.-C. Zhu, "Understanding human gaze communication by spatio-temporal graph reasoning," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2019, pp. 5724–5733.

[50] Y.-L. Li, L. Xu, X. Liu, X. Huang, Y. Xu, S. Wang, H.-S. Fang, Z. Ma, M. Chen, and C. Lu, "Pastanet: Toward human activity knowledge engine," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 382–391.

[51] T. Zhou, S. Qi, W. Wang, J. Shen, and S.-C. Zhu, "Cascaded parsing of human-object interaction recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021.

[52] H.-S. Fang, S. Xie, Y.-W. Tai, and C. Lu, "Rmpe: Regional multi-person pose estimation," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 2334–2343.

[53] S. Liu, J. Feng, C. Domokos, H. Xu, J. Huang, Z. Hu, and S. Yan, "Fashion parsing with weak color-category labels," *IEEE Trans. Multimedia*, vol. 16, no. 1, pp. 253–265, 2014.

[54] K. Yamaguchi, M. H. Kiapour, L. E. Ortiz, and T. L. Berg, "Parsing clothing in fashion photographs," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2012, pp. 3570–3577.

[55] W. Yang, P. Luo, and L. Lin, "Clothing co-parsing by joint image segmentation and labeling," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2014, pp. 3182–3189.

[56] N. Wang and H. Ai, "Who blocks who: Simultaneous clothing segmentation for grouping images," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2011, pp. 1535–1542.

[57] 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 *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2014, pp. 1971–1978.

[58] Y. Bo and C. C. Fowlkes, "Shape-based pedestrian parsing," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2011, pp. 2265–2272.

[59] J. Dong, Q. Chen, W. Xia, Z. Huang, and S. Yan, "A deformable mixture parsing model with parselets," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2013, pp. 3408–3415.

[60] J. Dong, Q. Chen, X. Shen, J. Yang, and S. Yan, "Towards unified human parsing and pose estimation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2014, pp. 843–850.

[61] K. Yamaguchi, M. Hadi Kiapour, and T. L. Berg, "Paper doll parsing: Retrieving similar styles to parse clothing items," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2013, pp. 3519–3526.

[62] H. Chen, Z. J. Xu, Z. Q. Liu, and S. C. Zhu, "Composite templates for cloth modeling and sketching," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2006, pp. 943–950.

[63] L. Ladicky, P. H. Torr, and A. Zisserman, "Human pose estimation using a joint pixel-wise and part-wise formulation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2013, pp. 3578–3585.

[64] S. Eslami and C. Williams, "A generative model for parts-based object segmentation," in *Proc. Advances Neural Inf. Process. Syst.*, 2012, pp. 100–107.

[65] I. Rauschert and R. T. Collins, "A generative model for simultaneous estimation of human body shape and pixel-level segmentation," in *Proc. Eur. Conf. Comput. Vis.*, 2012, pp. 704–717.

[66] 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 *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 1419–1427.

[67] 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 *Proc. IEEE Int. Conf. Comput. Vis.*, 2015, pp. 1386–1394.

[68] P. Luo, X. Wang, and X. Tang, "Pedestrian parsing via deep decompositional network," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2013, pp. 2648–2655.

[69] L.-C. Chen, Y. Yang, J. Wang, W. Xu, and A. L. Yuille, "Attention to scale: Scale-aware semantic image segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 3640–3649.

[70] 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 *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 648–663.

[71] 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*, 2017, pp. 7–15.

[72] S. Liu, C. Wang, R. Qian, H. Yu, R. Bao, and Y. Sun, "Surveillance video parsing with single frame supervision," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 413–421.

[73] J. Zhao, J. Li, Y. Cheng, T. Sim, S. Yan, and J. Feng, "Understanding humans in crowded scenes: Deep nested adversarial learning and a new benchmark for multi-human parsing," in *ACM MM*, 2018, pp. 792–800.

[74] Y. Luo, Z. Zheng, L. Zheng, T. Guan, J. Yu, and Y. Yang, "Macro-micro adversarial network for human parsing," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 418–434.

[75] 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 *The AAAI Conference on Artificial Intelligence*, 2018, pp. 7146–7153.

[76] K. Gong, Y. Gao, X. Liang, X. Shen, M. Wang, and L. Lin, "Graphonomy: Universal human parsing via graph transfer learning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 7450–7459.

[77] F. Xia, J. Zhu, P. Wang, and A. L. Yuille, "Pose-guided human parsing by an and/or graph using pose-context features," in *The AAAI Conference on Artificial Intelligence*, 2016, pp. 3632–3640.

[78] J. Li, C. Wang, H. Zhu, Y. Mao, H.-S. Fang, and C. Lu, "Crowdpose: Efficient crowded scenes pose estimation and a new benchmark," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 10 863–10 872.

[79] Y. Xiu, J. Li, H. Wang, Y. Fang, and C. Lu, "Pose flow: Efficient online pose tracking," in *Proc. of the British Machine Vision Conference*, 2018.

[80] Q. Li, A. Arnab, and P. H. Torr, "Holistic, instance-level human parsing," *arXiv preprint arXiv:1709.03612*, 2017.

[81] J. Li, J. Zhao, Y. Chen, S. Roy, S. Yan, J. Feng, and T. Sim, "Multi-human parsing machines," in *ACM MM*, 2018, pp. 45–53.

[82] 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.

[83] J. Zhao, J. Li, H. Liu, S. Yan, and J. Feng, "Fine-grained multi-human parsing," *Int. J. Comput. Vis.*, vol. 128, no. 8, pp. 2185–2203, 2020.

[84] Y. Freund and R. E. Schapire, "A decision-theoretic generalization of on-line learning and an application to boosting," *Journal of Computer and System Sciences*, pp. 119–139, 1997.

[85] C. Sutton, A. McCallum *et al.*, "An introduction to conditional random fields," *Foundations and Trends® in Machine Learning*, pp. 267–373, 2012.

[86] G. E. Hinton, "Training products of experts by minimizing contrastive divergence," *Neural computation*, pp. 1771–1800, 2002.

[87] K. Hornik, M. Stinchcombe, and H. White, "Multilayer feedforward networks are universal approximators," *Neural Networks*, pp. 359–366, 1989.

[88] S. Liang and R. Srikant, "Why deep neural networks for function approximation?" in *Proc. Int. Conf. Learn. Representations*, 2017.

[89] P. W. Battaglia, J. B. Hamrick, V. Bapst, A. Sanchez-Gonzalez, V. Zambaldi, M. Malinowski, A. Tacchetti, D. Raposo, A. Santoro, R. Faulkner *et al.*, "Relational inductive biases, deep learning, and graph networks," *arXiv preprint arXiv:1806.01261*, 2018.

[90] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 7132–7141.

[91] A. Veit and S. Belongie, "Convolutional networks with adaptive inference graphs," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 3–18.

[92] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. Advances Neural Inf. Process. Syst.*, 2017, pp. 5998–6008.
[93] X. Wang, R. Girshick, A. Gupta, and K. He, "Non-local neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 7794–7803.
[94] R. Hu, M. Rohrbach, and T. Darrell, "Segmentation from natural language expressions," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 108–124.
[95] X. Shi, Z. Chen, H. Wang, D.-Y. Yeung, W.-K. Wong, and W.-c. Woo, "Convolutional lstm network: A machine learning approach for precipitation nowcasting," in *Proc. Advances Neural Inf. Process. Syst.*, 2015, pp. 802–810.
[96] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 770–778.
[97] L.-C. Chen, G. Papandreou, F. Schroff, and H. Adam, "Rethinking atrous convolution for semantic image segmentation," *arXiv preprint arXiv:1706.05587*, 2017.
[98] K. Gong, X. Liang, Y. Li, Y. Chen, M. Yang, and L. Lin, "Instance-level human parsing via part grouping network," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 770–785.
[99] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 6230–6239.
[100] V. Badrinarayanan, A. Kendall, and R. Cipolla, "Segnet: A deep convolutional encoder-decoder architecture for image segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 12, pp. 2481–2495, 2017.
[101] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 801–818.
[102] P. Luc, C. Couprie, S. Chintala, and J. Verbeek, "Semantic segmentation using adversarial networks," in *NIPS-workshop*, 2016.
[103] M. M. Kalayeh, E. Basaran, M. Gokmen, M. E. Kamasak, and M. Shah, "Human semantic parsing for person re-identification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 1062–1071.
[104] X. Liu, M. Zhang, W. Liu, J. Song, and T. Mei, "BraidNet: Braiding semantics and details for accurate human parsing," in *ACM MM*, 2019, pp. 338–346.
[105] X. Liang, L. Lin, X. Shen, J. Feng, S. Yan, and E. P. Xing, "Interpretable structure-evolving lstm," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 1010–1019.
[106] 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 *Proc. IEEE Int. Conf. Comput. Vis.*, 2019, pp. 5218–5228.
[107] K. Dang and J. Yuan, "Location constrained pixel classifiers for image parsing with regular spatial layout," in *Proc. of the British Machine Vision Conference*, 2014.

Wenguan Wang received his Ph.D. degree from Beijing Institute of Technology in 2018. He is currently a postdoc researcher at ETH Zurich, Switzerland. From 2016 to 2018, he was a visiting Ph.D. student in University of California, Los Angeles. From 2018 to 2019, he was a senior scientist at Inception Institute of Artificial Intelligence, UAE. His current research interests include computer vision, image processing and deep learning.

Tianfei Zhou received his Ph.D. degree from Beijing Institute of Technology in 2017. He is currently a postdoc researcher at ETH Zurich, Switzerland. From 2018 to 2020, he was a research associate at Inception Institute of Artificial Intelligence, UAE. His current research interests include human-object interaction recognition, video object segmentation and deep learning.

Siyuan Qi received his B.Eng. degree in Computer Engineering from the University of Hong Kong in 2013. He received his M.S. and Ph.D degree in Computer Science from University of California, Los Angeles in 2015 and 2019, respectively. He is currently at Google, USA. His research interests include pattern recognition, machine learning and computer vision, with a focus on human activity recognition, scene understanding with compositional representations.

Jianbing Shen (M'11-SM'12) is currently acting as the Lead Scientist with the Inception Institute of Artificial Intelligence, UAE. He is also a Full Professor with the School of Computer Science, Beijing Institute of Technology. He has published about 100 journal and conference papers, and his Google scholar citations are more than 10000 times with H-index 50. He was awarded the Highly Cited Researcher by Web of Science in 2020. His research interests include computer vision, deep learning, autonomous driving, and medical image analysis. He is an Associate Editor of *IEEE TIP*, *IEEE TNNLS* and other journals.

Song-Chun Zhu received Ph.D. degree from Harvard University in 1996, and is Chair Professor jointly with Tsinghua University and Peking University, director of Institute for Artificial Intelligence, Peking University. He worked at Brown, Stanford, Ohio State, and UCLA before returning to China in 2020 to launch a non-profit organization – Beijing Institute for General Artificial Intelligence. He has published over 300 papers in computer vision, statistical modeling and learning, cognition, Language, robotics, and AI. He received a number of honors, including the Marr Prize in 2003, the Aggarwal prize from the Intl Association of Pattern Recognition in 2008, the Helmholtz Test-of-Time prize in 2013, twice Marr Prize honorary nominations in 1999 and 2007, a Sloan Fellowship, the US NSF Career Award, and ONR Young Investigator Award in 2001. He is a Fellow of IEEE since 2011. He serves as General co-Chair for CVPR 2012 and CVPR 2019.