Learning Semantic Neural Tree for Human Parsing

Ruyi Ji[1,2†], Dawei Du[3†], Libo Zhang[1,⋆], Longyin Wen[4], Yanjun Wu[1], Chen Zhao[1], Feiyue Huang[5], and Siwei Lyu[3]

[1] Institute of Software Chinese Academy of Sciences, China,
[2] University of Chinese Academy of Sciences, China,
[3] University at Albany, State University of New York, Albany, NY, USA,
[4] JD Finance America Corporation, Mountain View, CA, USA,
[5] Tencent Youtu Lab, China.

Abstract. In this paper, we design a novel semantic neural tree for human parsing, which uses a tree architecture to encode physiological structure of human body, and design a coarse to fine process in a cascade manner to generate accurate results. Specifically, the semantic neural tree is designed to segment human regions into multiple semantic sub-regions (*e.g.*, face, arms, and legs) in a hierarchical way using a new designed attention routing module. Meanwhile, we introduce the semantic aggregation module to combine multiple hierarchical features to exploit more context information for better performance. Our semantic neural tree can be trained in an end-to-end fashion by standard stochastic gradient descent (SGD) with back-propagation. Several experiments conducted on four challenging datasets for both single and multiple human parsing, *i.e.*, LIP, PASCAL-Person-Part, CIHP and MHP-v2, demonstrate the effectiveness of the proposed method. Code can be found at `https://isrc.iscas.ac.cn/gitlab/research/sematree`.

Keywords: human parsing·semantic neural tree·semantic segmentation

1 Introduction

Human parsing aims to recognize each semantic part, *e.g.*, arms, legs and clothes, which is one of the most fundamental and critical problems in analyzing human with various applications, such as video surveillance, human-computer interaction, and person re-identification. With the development of convolutional neural networks (CNN) on semantic segmentation task, human parsing has obtained significant accuracy improvement recently. Most of previous algorithms [40, 9, 3, 28] attempt to assign each pixel with the predefined semantic labels, such as *arm* and *leg*. However, each semantic label is considered independently, which fails to consider context relations among different semantic labels, *e.g.*, the *upper-body* region is formed by the *torso*, *upper-arms* and *lower-arms* regions, see Fig. 1.

⋆ Libo Zhang is the corresponding author(libo@iscas.ac.cn). † denotes equal contribution.

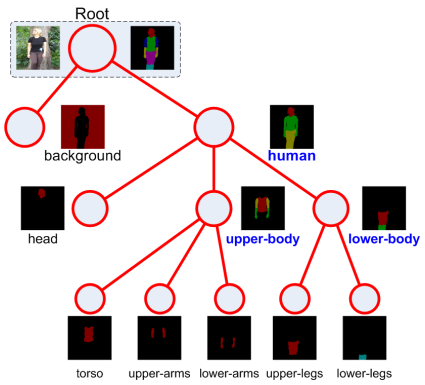

Fig. 1. Category hierarchy used in the PASCAL-Person-Part dataset [4].

Thus, exploiting the intrinsic physical structure of human body is an effective way to improve the segmentation accuracy.

Inspired from human perception [17], we argue that it is reasonable to use the hierarchical structure network to exploit discriminative features of human body to solve the human parsing task. Thus, we design a semantic neural tree network to encode the physical structure of human body, and design a coarse to fine process in a cascade manner. The coarse to fine process in a hierarchical design is helpful to improve the performance of human parsing. As an example in Fig. 1, we introduce a virtual category *upper-body*, and first distinguish the *upper-body* from the *head* and *lower-body* pixels. After that, we segment the *torso*, *upper-arms*, and *lower-arms* regions from the segmented *upper-body* region. Thus the hierarchical design in the cascade manner can generate more accurate results.

In this paper, we design a novel semantic neural tree (SNT) for human parsing, which uses a tree architecture to encode physiological structure of human body and design a coarse to fine process in a cascade manner. According to the topology structure of annotations in different datasets, we can design different tree architecture in a similar spirit. For the leaf node of each path in the tree, our goal is to distinguish just a few categories. In general, the proposed semantic neural tree consists of four components, *i.e.*, the backbone network for feature extraction, attention routing modules for sub-category partition, semantic aggregation modules for discriminative feature representation and prediction modules for generating parsing results, laid in several levels. That is, we segment human regions into multiple semantic sub-regions in a hierarchical way using the attention routing module. After that, we introduce the semantic aggregation module to combine multiple hierarchical features to exploit more context information. We generate the parsing result by aggregating the discriminative feature maps from each leaf node. Our SNT is trained in an end-to-end fashion using the standard stochastic gradient descent (SGD) with back-propagation [19].

Several experiments are conducted on four challenging datasets, *i.e.*, LIP [22], Pascal-Person-Part [4], CIHP [9] and MHP-v2 [41], demonstrating that our SNT

method achieves favorable performance against the state-of-the-art methods for both single and multiple human parsing. Meanwhile, we also carry out ablation experiments to validate the effectiveness of the components in our SNT. The main contributions are summarized as follows, (1)We propose a semantic neural tree for human parsing, which integrates the physiological structure of human body into a tree architecture, and design a coarse to fine process in a cascade manner; (2)We introduce the semantic aggregation module to combine multiple hierarchical features to exploit more context information; (3)The experimental results on several challenging single and multiple human parsing datasets demonstrate that the proposed method achieves favorable performance against the state-of-the-art methods.

2 Related Work

Semantic segmentation. Towards accurate scene understanding, many researchers [40, 26, 2, 3, 1] propose semantic segmentation methods based on the fully convolutional network (FCN) [29]. Zhao *et al.* [40] propose the pyramid scene parsing network (PSPNet) to capture the capability of global context information by different-region based context aggregation. In [26], the multi-path refinement network is developed to extract all the information available along the down-sampling process to enable high-resolution prediction using long-range residual connections. Besides, Chen *et al.* [2] introduce atrous spatial pyramid pooling (ASPP) to segment objects at multiple scales accurately. Improved from [2], they apply the depth-wise separable convolution to both ASPP and decoder modules to refine the segmentation results especially along object boundaries [3]. Recently, the meta-learning technique is applied in image prediction focused on the tasks of scene parsing, person-part segmentation, and semantic image segmentation, resulting in better performance [1]. However, these semantic segmentation methods are constructed without considering the relations among semantic sub-categories, leading to limited performance for human parsing with fine-grained sub-categories.

Human parsing. Furthermore, human parsing can be regarded as a fine-grained semantic segmentation task. To adapt to the human parsing task, more useful modules are proposed and combined in the semantic segmentation methods. Ruan *et al.* [28] improve the PSPNet [40] by using the global context embedding module for multi-scale context information. Zhao *et al.* [41] employ three Generative Adversarial Network-like networks to perform semantic saliency prediction, instance-agnostic parsing and instance-aware clustering respectively. However, the aforementioned methods prefer to construct complex network for more discriminative representation, but consider little about semantic structure of human body when designing the network.

The semantic structure information is essential in human parsing. Gong *et al.* [9] consider instance-aware edge detection to group semantic parts into distinct person instances. Liang *et al.* [22] propose a novel joint human parsing and pose estimation network, which imposes human pose structures into the pars-

ing results without resorting to extra supervision. In [8], the hierarchical graph transfer learning is incorporated upon the parsing network to encode the underlying label semantic structures and propagate relevant semantic information. Different from them without exploring human hierarchy, we take full use of the category label hierarchy and propose a new tree architecture to learn semantic regions in a coarse to fine process.

It is worth mentioning that some previous methods [35, 24, 33, 34] use body physical structure information to improve the human parsing accuracy. Wang *et al.* [35] introduce hierarchical poselets to represent the rigid parts, covering large portions of the human body. Liang *et al.* [24] formulate the human parsing task as the active template regression problem, which uses the linear combination of the learned mask templates to represent each body item. The aforementioned two methods rely on keypoints detection to exploit the intrinsic structure information of human body, which brings extra computational overhead and relies on additional keypoints annotations in the training phase. The method [33] uses tree-like topology in network structure to fuse the information from three levels, *i.e.*, down-top, top-down, which assembles information from three inference processes over the hierarchy. Wang *et al.* [34] explore three categories of part relations, *i.e.*, decomposition, composition, and dependency, to simultaneously exploit the representation learning capacity of deep graph networks and the hierarchical human structures. In contrast to the method [34] focusing on the particular relations between nodes, our method designs the neural tree to encode the physical structure of human body in a coarse-to-fine manner. Meanwhile, different branches in the same hierarchy focus on different subregions of human body, and different hierarchies focus on human body with different receptive fields.

Neural tree. The decision tree (DT) is an effective model and widely applied in machine learning tasks. As the inherent of the interpretability, it is usually regarded as an auxiliary tool to insight into the mechanism of neural network. However, the simplicity of identity function used in these methods means that input data is never transformed and thus each path from root to leaf node on the tree does not perform representation learning, limiting their performance. To integrate non-linear transformations into DTs, Kontschieder *et al.* [18] propose the stochastic and differentiable decision tree model based neural decision forest. Similarly, Xiao *et al.* [38] develop a neural decision tree with a multi-layer perceptron network at the root transformer. In contrast, our model focuses on the "topology structure" of annotations (see Fig. 1). That is, the proposed model have a flexible semantic topology depending on certain dataset.

3 Methodology

The goal of the proposed Semantic Neural Tree (SNT) method is to classify local parts of human along the path from root to leaf, and then fuse the feature maps before each leaf node to form the global representation for parsing prediction. We depart each sample $x \in X$ with the parsing label $y \in Y$. Notably, our

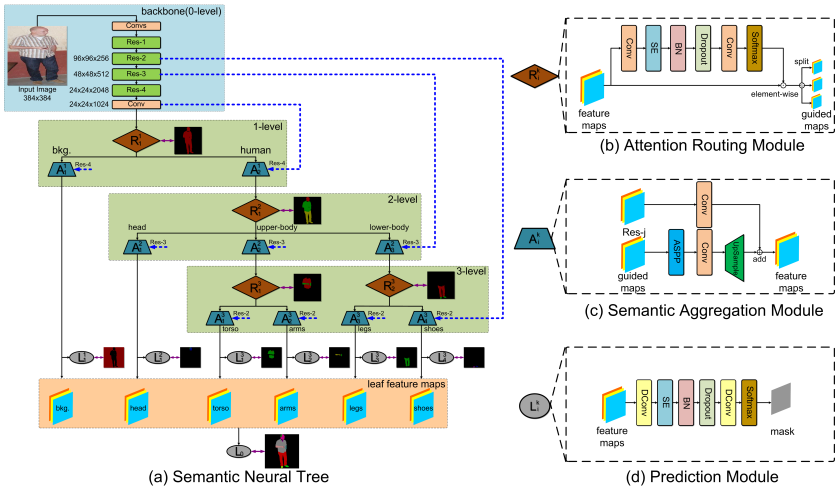

Fig. 2. The tree architecture of our SNT model used on the LIP dataset [22]. The blue dashed lines indicate that the semantic aggregation module in each level merges the features from different layers in the backbone. The purple double arrows denote the supervision for the attention routing and prediction modules. Best view in color.

model is not a full binary tree, because the topology of model is determined by the semantics of dataset. Based on our tree architecture, we group the parsing labels into category label hierarchy. For example, as shown in Fig. 2(a), the virtual category label *head* consists of several child category labels *face*, *hair* and *hat* in the LIP dataset [22].

Our model consists of four modules, the backbone network, the attention routing module, the semantic aggregation module, and the prediction module. Specifically, the backbone network is used to extract features in the proposed method. The attention routing module is designed to generate masks to determine the root-to-leaf computation paths. In this way, different branches in the same hierarchy of the tree focus on different subregions of human body, and different hierarchies in the tree network focus on human body with different receptive fields. Meanwhile, the semantic aggregation module integrates ASPP and SE modules to enforce the network to exploit discriminative features. After that, the prediction module is used to generate the parsing results for each category. We organize these four modules in a tree architecture and solve the parsing task in a coarse-to-fine process for accurate results.

3.1 Architecture

Backbone network. Similar to the previous works, we rely on residual blocks of ResNet-101 network [13] to extract discriminative features of human in each sub-category. Our SNT can also work on other pre-trained networks, such as DenseNet [15] and Inception [32]. Specifically, we remove the global average

pooling and fully connected layers from the network and use the truncated ResNet-101 network [13], *i.e.*, Res-j, ($j = 1, 2, 3, 4$), as the backbone. Meanwhile, followed by the backbone, we add one convolutional layer with the kernel size 1×1 and stride size 1 to reduce the channels of feature maps Res-4. Notably, as shown in Fig. 2, we employ multi-scale feature representation as a powerful tool to improve the ResNet-101 backbone in the dense prediction task with highly localized discriminative regions in fine-grained categories.

Attention routing module. After the backbone network, we need to solve how to split the tree structure. Given the sample x, in each level of the tree architecture, we employ the attention routing module to split the higher-level category labels and output the corresponding intermediate masks. That is, the i-th attention routing module at the k-th level R_i^k is fed with the feature maps $\phi_i^{k-1}(x)$ at the $(k-1)$-th level. To this end, we supervise R_i^k based on the labels of pre-set virtual categories.

As shown in Fig. 2(b), the attention routing module starts from one convolutional layer with the kernel size 1×1 and one Squeeze-and-Excitation (SE) layer [14]. Thus we can reduce the computational complexity and enforce the model to pay more attention to discriminative regions. After that, we use one dropout layer with the drop rate 0.5, one convolutional layer with the kernel size 1×1 and one softmax layer to output the mask of the pixel-level human parts $\Psi_i^k(x) = \{\psi_1^k(x), \cdots, \psi_I^k(x)\}$ such that $\psi_i^k(x) \in [0, 1]$. Notably, the channels of the mask consists of foreground channels and background channel, where I denotes the channel number of $\Psi_i^k(x)$. The foreground channels denote the sub-category labels at node i while background channel is defined as the other labels excluded from the sub-category labels at node i. With supervision on the masks, we can guide and split the feature maps at the k-th level into several semantic sub-categories, *i.e.*, $\Phi_i^k(x) = \{\phi_1^k(x), \cdots, \phi_I^k(x)\}$.

Semantic aggregation module. Followed by the attention routing module R_i^k, our goal is to extract discriminative feature representation for sub-categories. To this end, multi-scale feature representation is an important and effective strategy, *e.g.*, skip-connections in the U-Net architecture [3]. Besides, the convolution with stride larger than one and the pooling operations will shrink feature maps, resulting in information loss in details such as the edge or small parts.

To alleviate these issues, we introduce the semantic aggregation module A_i^k to deal with the feature maps $\phi_i^k(x)$. Specifically, we first adapt atrous spatial pyramid pooling (ASPP) [2] to concatenate the features from multiple atrous convolutional layers with different dilation rates arranged in parallel. Specifically, the ASPP module is built to deal with the guided feature maps after the semantic router with dilation rates $[1, 6, 12, 18]$ to form multi-scale features. To aggregate multi-scale feature, we also use the upsampling layer to increase the spatial size of feature while halve the number of channels. After that, we use the addition operation to fuse the multi-scale features from the ASPP module and the residual features of the backbone Res-j at the j-th stage (see Fig. 2(c)). Thus we can learn more discriminative feature maps $\hat{\phi}_i^k(x)$ for prediction.

Prediction module. Based on the feature maps after semantic aggregation $\hat{\phi}_i^k(x)$, we use the prediction modules L_i^k in different levels to generate the parsing result for each sub-category. As shown in Fig. 2(d), the prediction module includes one deformable convolutional layer [43] with the kernel size 3×3, one SE layer [14], one batch normalization layer, one dropout layer with drop rate 0.5 and another deformable convolutional layer [43] with the kernel size 3×3. Finally, the softmax layer is used to output an estimate for conditional distribution for each pixel. For each leaf node at the k-th level, we can predict the local part parsing result $\varphi_i^k(x)$.

Moreover, we combine all the feature maps of each leaf node $\hat{\phi}_i^k(x)$. Specifically, we remove the background channel in every leaf feature map and then concatenate the rest foreground channels, *i.e.*, background, head, torso, arms, legs and shoes in Fig. 2(a), such that the overall number of channels is equal to the number of categories. Thus we can predict the final parsing result $\mathcal{P}(x)$ by using the prediction module L_0.

3.2 Loss function

Class imbalance is an important issue that results in reduced performance easily. A common solution is to perform the hard negative mining strategy that samples hard examples or more other sampling/reweighing schemes during training phase. Since we aggregate several sub-categories into one virtual category in coarse levels, more severe class imbalance may exist in our hierarchical tree model. To deal with this issue, we adopt a simple category re-weighting strategy. Specifically, based on the ground-truth mask, we calculate percentage of pixels belonging to each category in every batch. Without doubt, the background is overwhelming compared with other categories. Therefore we consider the loss of pixels belonging to each category using the corresponding weight as $\mathcal{W}_{\mathcal{X}^j} = 1 - \sum_i \mathcal{C}_{\mathcal{X}_i^j} / \sum_j \sum_i \mathcal{C}_{\mathcal{X}_i^j}$, where $\mathcal{C}_{\mathcal{X}_i^j}$ indicates the i-th pixel belongs to the j-th category in $\mathcal{X}$ module in the current batch. $\mathcal{X}$ modules consist of the attention routing module (R), leaf node parsing (L) and final parsing (L_0). Based on the weights, we use three loss terms on the attention routing module, each leaf node, and the final output after prediction modules to train the whole network in an end-to-end manner, which is computed as

$$\begin{aligned}\mathcal{L} = &\sum_i \sum_k \mathcal{L}_{R_i^k}(\Psi_i^k(x), \bar{y}_i^k, \mathcal{W}_R) + \omega_1 \cdot \sum_i \sum_k \mathcal{L}_{L_i^k}(\varphi_i^k(x), \dot{y}_i^k, \mathcal{W}_L) \\ &+ \omega_2 \cdot \mathcal{L}_{L_0}(\mathcal{P}(x), y^*, \mathcal{W}_{L_0}),\end{aligned} \tag{1}$$

where $\mathcal{L}_{R_i^k}(\cdot,\cdot,\cdot)$ denotes the re-weighted cross-entropy loss between the masks $\Psi_i^k(x)$ generated by the attention routing module R_i^k and the corresponding ground-truth $\bar{y}_i^k$ at the k-th level. $\mathcal{L}_{L_i^k}(\cdot,\cdot,\cdot)$ denotes the re-weighted cross-entropy loss between the output map $\varphi_i^k(x)$ by the leaf node and the corresponding ground-truth map $\dot{y}_i^k$ at the k-th level. $\mathcal{L}_{L_0}(\cdot,\cdot,\cdot)$ denotes the re-weighted cross-entropy loss between the final parsing result $\mathcal{P}(x)$ and the global parsing

label y^*. The factors ω_1 and ω_2 are used to balance the attention routing module, leaf node parsing and final parsing. $\mathcal{W}_R$, $\mathcal{W}_L$ and $\mathcal{W}_P$ are category weights on router module, leaf node and global parsing, respectively. It is worth mentioning that the channel number of $\bar{y}_i^k$ is equal to the number of sub-category labels of node i at the k-th level, and the channel number of y^* is equal to the total number of labels.

3.3 Handling multiple human parsing

To handle multiple human parsing, we integrate our method with the off-the-shelf instance segmentation framework, as similar as that in [28]. Specifically, we first employ the Mask R-CNN [12] pre-trained on MS-COCO dataset [27] to segment human instances from images. Then, we train three SNT sub-models to obtain global and local human parsing results with different size of input images, *i.e.*, one global sub-model and two local sub-models. The global sub-model is trained on the whole images without distinguishing each instance; while the other two local sub-models are input by segmented instance patches from Mask R-CNN [12] and ground-truth respectively. Notably, we use the same architecture for the three sub-models. Finally, both the global and local results from these sub-models are combined to output multiple human parsing results by late fusion. That is, we concatenate the feature maps before leaf node on each sub-branches in our network. Followed by the prediction module, we can estimate the category for each pixel under the supervision of cross-entropy loss function.

4 Experiment

Following the previous works [28, 9, 42, 8], we compare our method with other state-of-the-art methods on the validation set of two single human parsing datasets (*i.e.*, LIP [22] and Pascal-Person-Part [4]) and two multiple human parsing datasets (*i.e.*, CIHP [9] and MHP-v2 [41]). Different evaluation datasets have different definitions of the topology of human body. Note that the physical structure of human body is intrinsic, such as head, arms, and legs. Thus, the annotations of human body can be easily obtained based on the uniform definition of the tree topology of human body.

Implementation Details. We implement the proposed framework in PyTorch. All models are trained on a workstation with a 3.26 GHz Intel processor, 32 GB memory, and one Nvidia V100 GPU. Following the previous works, we adopt the ResNet-101 [13] that is pre-trained on the ImageNet dataset [5] as the backbone network. For a fair comparison, we set input size of images 384×384 for single person parsing while 473×473 for multiple person parsing. For data argumentation, we adopt the strategy of random scaling (from 0.5 to 1.5), random rotation, random cropping and left-right flipping the training data. We use the SGD algorithm to train the network with 0.9 momentum, and 0.0005 weight decay. The learning rate is initialized to 0.001 and adjusted by exponential learning

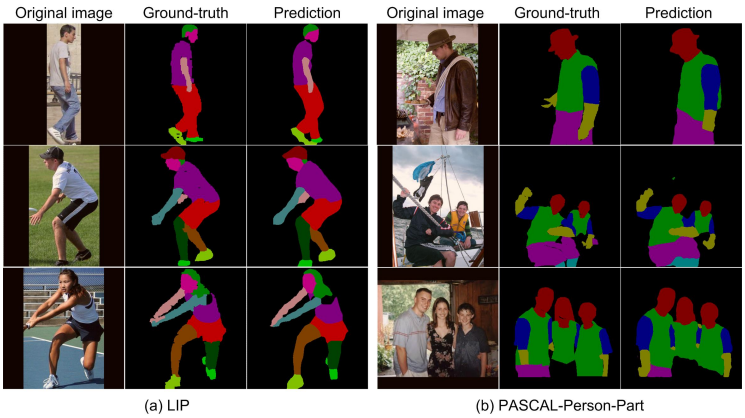

Fig. 3. Some visualized examples of single human parsing.

rate decay policy (gamma is 0.9). Notably, the warming up policy is applied for training. That is, we use the learning rate of 0.0001 to warm up the model in the first 10 epochs, and then increase learning rate up to 0.001 linearly. The model is optimized in 200 epochs. In the loss function (1), the weights ω_1 and ω_2 are set as 1.5 and 2.0 empirically.

Metrics. First, we employ the mean IoU metric (mIOU) to evaluate the global-level predictions in single human parsing datasets (*i.e.*, LIP [22] and Pascal-Person Part [4]). Then, we use three metrics (*i.e.*, AP^r, AP^p and PCP) to evaluate the instance-level predictions in multiple human parsing. The AP^r score denotes the area under the precision-recall curve based on the limitation of different IoU thresholds (*e.g.*, 0.5, 0.6, 0.7) [11]. PCP elaborates how many body parts are correctly predicted of a certain person [20]. AP^p computes the pixel-level IoU of semantic part categories within a person. Similar to the previous works, we use the metrics of mIoU and AP^r to evaluate the performance on the CIHP dataset [9] while PCP and AP^p to evaluate the performance on the MHP-v2 dataset [41]. AP^r_m denotes the mean value.

4.1 Single Human Parsing

We compare the performance of single human parsing of our proposed method with other state-of-the-arts on the LIP [22] and Pascal-Person-Part [4] datasets. The qualitative human parsing results are visualized in Fig. 3.

Evaluation on LIP Dataset. The LIP dataset defines 6 body parts and 13 clothes categories, including $50,462$ images with pixel-level annotations. $30,462$ training and $10,000$ validation images are provided with publicly available annotations. As shown in Fig. 2, we construct the tree architecture in 3-level. As presented in Table 1, we can conclude that our method achieves the best performance in terms of all the three metrics. Since semantic segmentation methods

Table 1. The evaluation results on the validation set of LIP [22].

Method	pixel acc.	mean acc.	mIoU
Attention+SSL [10]	-	-	44.73
DeepLab [2]	84.09	55.62	44.80
MMAN [30]	-	-	46.81
SS-NAN [39]	87.60	56.00	47.92
MuLA [31]	**88.50**	60.50	49.30
PSPNet [40]	86.23	61.33	50.56
JPPNet [21]	86.39	62.32	51.37
CE2P [28]	-	-	52.56
CE2P(w/ flip) [28]	87.37	63.20	53.10
Ours	88.10	**70.41**	**54.86**

Table 2. The evaluation results on the validation set of LIP [22] in each category.

Method	bkg.	hat	hair	glove	glasses	u-clothes	dress	coat	socks	pants	
Attention+SSL [10]	84.6	59.8	67.3	29.0	21.6	65.3	29.5	51.9	38.5	68.0	
DeepLab [2]	84.1	59.8	66.2	28.8	23.9	65.0	33.7	52.9	37.7	68.0	
PSPNet [40]	86.1	63.5	68.0	39.1	23.8	68.1	31.7	56.2	44.5	72.7	
MMAN [30]	84.8	57.7	65.6	30.1	20.0	64.2	28.4	52.0	41.5	71.0	
JPPNet [21]	86.3	63.6	70.2	36.2	23.5	68.2	31.4	55.7	44.6	72.2	
CE2P [28]	87.4	64.6	72.1	38.4	32.2	68.9	32.2	55.6	48.8	73.5	
Ours	**88.2**	**66.9**	**72.2**	**42.7**	**32.3**	**70.1**	**35.6**	**57.5**	**48.9**	**75.2**	
Method	j-suit	scarf	skirt	face	l-arm	r-arm	l-leg	r-leg	l-shoe	r-shoe	mIoU
Attention+SSL [10]	24.5	14.9	24.3	71.0	52.6	55.8	40.2	38.8	28.1	29.0	44.7
DeepLab [2]	26.1	17.4	25.2	70.0	50.4	53.9	39.4	38.3	27.0	28.4	44.8
PSPNet [40]	28.7	15.7	25.7	70.8	59.7	62.3	54.9	54.5	42.3	42.9	50.6
MMAN [30]	23.6	9.7	23.2	69.5	55.3	58.1	51.9	52.2	38.6	39.0	46.8
JPPNet [21]	28.4	18.8	25.1	73.4	62.0	63.9	58.2	58.0	44.0	44.1	51.4
CE2P [28]	27.2	13.8	22.7	74.9	64.0	65.9	59.7	58.0	45.7	45.6	52.6
Ours	**33.4**	**21.4**	**27.4**	**74.9**	**66.8**	**68.1**	**60.3**	**59.8**	**47.6**	**48.1**	**54.8**

(*e.g.*, DeepLab [2] and PSPNet [40]) consider little about fine-grained classification in the human parsing task, they perform not well. Moreover, the CE2P method [28] improves PSPNet [40] by adding the context embedding branch, achieving 53.10 mIOU score. Our method exceeds CE2P by 1.76% in terms of mIOU score. It indicates that our method can learn discriminative representation of each sub-category for human parsing. Moreover, as shown in Table 2, our method obtain the best mIOU score in each sub-category. Notably, our method achieves considerable accuracy improvement compared with the other methods in some ambiguous sub-categories, *e.g.*, *glove*, *j-suit*, and *shoe*.

Evaluation on Pascal-Person-Part Dataset. The PASCAL-Person-Part dataset [4] is originally from the PASCAL VOC-2010 dataset [6], and then extended for human parsing with 6 coarse body part labels. It consists of $1,716$ training images and $1,817$ testing images ($3,533$ images in total). As shown in Fig. 1, we construct the tree architecture in 3-level. Specifically, the virtual category

Table 3. The evaluation results on the validation set of Pascal-Person-Part [4].

Method	head	torso	u-arms	l-arms	u-legs	l-legs	bkg.	mIoU
HAZN [36]	80.79	**80.76**	45.65	43.11	41.21	37.74	93.78	57.54
Attention+SSL [10]	83.26	62.40	47.80	45.58	42.32	39.48	94.68	59.36
Graph LSTM [25]	82.69	62.68	46.88	47.71	45.66	40.93	94.59	60.16
SE LSTM [23]	82.89	67.15	51.42	48.72	51.72	45.91	97.18	63.57
Part FCN [37]	85.50	67.87	54.72	54.30	48.25	44.76	95.32	64.39
DeepLab [2]	-	-	-	-	-	-	-	64.94
MuLA [31]	-	-	-	-	-	-	-	65.10
SAN [16]	86.12	73.49	59.20	56.20	51.39	49.58	96.01	64.72
WSHP [7]	87.15	72.28	57.07	56.21	52.43	50.36	97.72	67.60
DeepLab v3+ [3]	-	-	-	-	-	-	-	67.84
PGN [9]	**90.89**	75.12	55.83	64.61	55.42	41.57	95.33	68.40
Graphonomy [8]	-	-	-	-	-	-	-	69.12
Compositional Fusion [33]	88.02	72.01	**64.31**	63.52	55.61	**54.96**	96.02	70.76
DPC [1]	88.81	74.54	63.85	63.73	57.24	54.55	96.66	71.34
Ours	89.01	74.63	62.90	**64.70**	**57.53**	54.62	**97.74**	**71.59**

human consists of three sub-categories, *i.e.*, *head*, *upper-body* including torso, upper-arms and lower-arms and *lower-body* including upper-legs and lower-legs. We report the performance on the Pascal-Person-Part dataset in Table 3. Similar to the trend in the LIP dataset [22], the semantic segmentation methods, *e.g.*, DeepLab [2] and DeepLab v3+ [3], perform inferior mIoU score, *i.e.*, less than 68.00. Moreover, the Graphonomy method [8] learns and propagates compact high-level graph representation among the labels within one dataset, resulting in better 69.12 mIoU score. Besides, DPC [1] achieves better performance with 71.34 mIoU score. This is because it employs meta-learning to search optimal efficient multi-scale network for human parsing. Our SNT method obtains the best overall mIoU score of 71.59 and best mIoU scores in terms of *l-arms* and *u-legs* among all the compared methods, which indicates the effectiveness of our proposed tree network.

4.2 Multiple Human Parsing

Furthermore, we evaluate the proposed method on two large-scale multiple human parsing datasets, *i.e.*, CIHP [9] and MHP-v2 [41]. For a fair comparison, we apply same Mask R-CNN model to output instance segmentation masks. Then, we use the global parsing and two local parsing models for human parsing as in [28]. Following the [28], final results are obtained by fusing the results from three branch models with a refinement process. Some visual results are shown in Fig. 4, which indicates that our method can also generate precise and fine-grained results in multiple human parsing scenes.

Evaluation on CIHP Dataset. The CIHP dataset [9] is the largest multi-person human parsing dataset with $38,280$ diverse human images, *i.e.*, $28,280$ training, $5,000$ validation and $5,000$ test images. We use the same topology

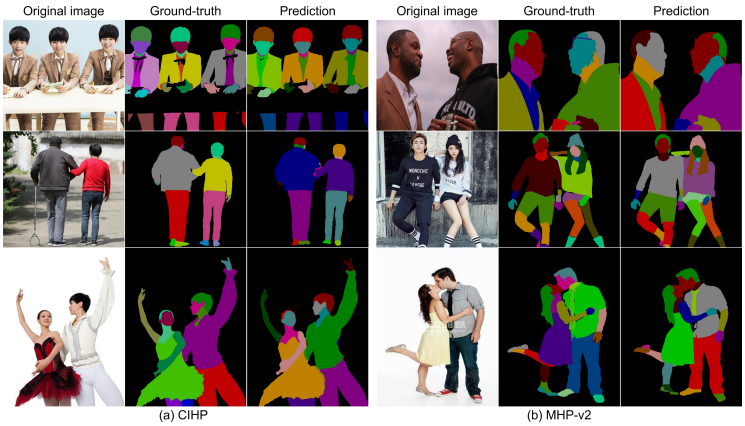

Fig. 4. Some visualized examples of multiple human parsing.

Table 4. The evaluation results on the validation set of CIHP [9].

Method	mIoU	$AP^r_{0.5}$	$AP^r_{0.6}$	$AP^r_{0.7}$	AP^r_m
PGN [9]	55.89	35.80	28.60	20.50	33.60
M-CE2P [28]	59.50	48.69	40.13	29.74	42.83
Ours	**60.87**	**49.27**	**41.98**	**33.00**	**43.96**

(*i.e.*, 3-level tree structure as shown in Fig. 2) in the LIP dataset [22] to perform human parsing because the two datasets share the same sub-category semantic annotations. As shown in Table 4, our method outperforms other compared methods (*i.e.*, PGN [9] and M-CE2P [28]), achieving AP^r_m score of 43.96. It is worth mentioning that SNT outperforms M-CE2P [28] in terms of $AP^r_{0.7}$ score by considerable improvement, *i.e.*, 29.74 vs. 33.00. It indicates that our method facilitates improving the segmentation accuracy of human instances.

Evaluation on MHP-v2 Dataset. The MHP-v2 dataset [41] includes 25,403 annotated images with 58 fine-grained semantic category labels. Since this dataset has more labels than the LIP dataset [22], we construct the tree architecture in 5-level. As shown in Table 5, the semantic segmentation method Mask R-CNN [12] has the worst performance compared to other methods. NAN [42] achieves the AP^p_m score of 42.77, but much inferior performance in both $PCP_{0.5}$ and $AP^p_{0.5}$ scores. Our method achieves comparable state-of-the-art performance with M-CE2P [28] in terms of three metrics. It indicates that the coarse to fine process in a hierarchical design can facilitate improving the accuracy.

4.3 Ablation study

We study the influence of some important parameters and components of our SNT method as follows. The experiment is conducted on the LIP dataset [22].

Height of the tree. The height of the tree k indicates the complexity of the network. To explore the optimal height, we design five variants with different

Table 5. The performance on the validation set of MHP-v2 [41].

Method	$PCP_{0.5}$	$AP^p_{0.5}$	AP^p_m
Mask R-CNN [12]	25.12	14.50	-
MH-Parser [23]	26.91	18.05	-
NAN [42]	34.37	24.87	42.77
M-CE2P [28]	**43.77**	34.47	42.70
Ours	43.50	**34.76**	**43.03**

Table 6. Effect of the height of the tree on the LIP dataset [22].

height of the tree	pixel acc. (%)	mean acc. (%)	mIoU (%)
0	84.81	57.12	46.34
1	86.84	64.03	52.15
2	87.42	65.58	53.32
3	**88.10**	**70.41**	**54.86**
4	86.92	64.34	51.42

heights of the tree, see Fig. 2(a). If the height is equal to 0, only the ResNet-101 backbone is used for human parsing. As presented in Table 6, we can observe there is a sharp decline in mean accuracy and mIoU score. We find that our method with 3-level achieves the best performance, *i.e.*, 54.86% mIoU score. The performance of our method sharply drops, using deeper or shallower tree architectures. If we set the tree height $k \leq 2$, there are limited number of parameters in our model, not enough to represent the significant variations of human body. Meanwhile, if we set $k = 4$, too many parameters with limited number of training data cause overfitting.

Effectiveness of prediction module. To analyze prediction module in the proposed network, we construct a variant of our method, *i.e.*, "ours w/o pred". As shown in Fig. 2(d), the "ours w/o pred" method indicates that we combine the prediction results of each leaf node for final parsing result without the prediction module. If we do not use the prediction module to generate the final parsing result, we can observe a sharp decrease in mIoU score, *i.e.*, 50.02 vs. 54.86. It is essential to achieve accurate parsing result based on the context information among every sub-categories.

Effectiveness of semantic aggregation module. To verify the effectiveness of the semantic aggregation module, we construct the "ours w/o skip" method, which indicates that we do not combine the residual blocks from the backbone in semantic aggregation (see Fig. 2(c)). According to Table 7, we can conclude that the skip-connection from the backbone (see the dashed blue lines in Fig. 2(a)) can bring 1.54% mIoU improvement. This is because the skip-connection in our network can exploit multi-scale representation for sub-categories.

Effectiveness of attention routing module. To study the effect of the attention routing module, the "ours w/o mask" indicates that we further remove the attention mask in the attention routing module from the "ours w/o skip" method

Table 7. Variants of the SNT method on the LIP dataset [22].

variant	pixel acc. (%)	mean acc. (%)	mIoU (%)
Ours w/o mask	86.84	64.03	52.15
Ours w/o skip	87.42	65.58	53.32
Ours w/o pred	85.34	63.22	50.02
Ours w/o reweight	87.92	66.42	54.73
Ours w/o warming up	87.95	70.21	54.75
Ours	**88.10**	**70.41**	**54.86**

(see Fig. 2(b)). That is, we directly split the feature maps into several semantic maps for the next level. As presented in Table 7, the "ours w/o skip" method achieves 1.17% improvement in mIoU score compared with the "ours w/o mask" method. It demonstrates the attention mask can enforce the tree network focus on discriminative representation for specific sub-category semantic information.
Effectiveness of reweighting strategy. To verify the effect of the reweighting strategy, we construct a variant by removing the reweighting strategy. That is, we train the network with equal weights of different categories. As presented in Table 7, the "ours w/o reweight" method drops 3.99% mean acc. score compared with our method, which demonstrates the effectiveness of the reweighting strategy.
Effectiveness of warming up strategy. To demonstrate the influence of the warming up policy, we construct a variant of the proposed method by removing the warming up policy, denoted as "ours w/o warming up". Specifically, the average accuracy of the "ours w/o warming up" method on the LIP dataset drops 0.2% compared to the proposed method, which demonstrates the effectiveness of the warming up policy.

5 Conclusion

In this paper, we propose a novel semantic tree network for human parsing. Our method can encode physiological structure of human body and segment multiple semantic sub-regions in a hierarchical way. Extensive experiments on four challenging single and multiple human parsing datasets indicate the effectiveness of the proposed semantic tree structure. Our method can learn discriminative feature representation and exploit more context information for sub-categories effectively. For future work, we plan to optimize the tree architecture for better performance by neural architecture search techniques.

Acknowledgements

This work was supported by the Key Research Program of Frontier Sciences, CAS, Grant No. ZDBS-LY-JSC038, the National Natural Science Foundation of China, Grant No. 61807033. Libo Zhang was supported by Youth Innovation Promotion Association, CAS (2020111), and Outstanding Youth Scientist Project of ISCAS.

References

1. Chen, L., Collins, M.D., Zhu, Y., Papandreou, G., Zoph, B., Schroff, F., Adam, H., Shlens, J.: Searching for efficient multi-scale architectures for dense image prediction. In: NeurIPS. vol. abs/1809.04184 (2018)
2. Chen, L., Papandreou, G., Kokkinos, I., Murphy, K., Yuille, A.L.: Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. TPAMI **40**(4), 834–848 (2018)
3. Chen, L., Zhu, Y., Papandreou, G., Schroff, F., Adam, H.: Encoder-decoder with atrous separable convolution for semantic image segmentation. In: ECCV. pp. 833–851 (2018)
4. Chen, X., Mottaghi, R., Liu, X., Fidler, S., Urtasun, R., Yuille, A.L.: Detect what you can: Detecting and representing objects using holistic models and body parts. In: CVPR. pp. 1979–1986 (2014)
5. Deng, J., Dong, W., Socher, R., Li, L., Li, K., Li, F.: Imagenet: A large-scale hierarchical image database. In: CVPR. pp. 248–255 (2009)
6. Everingham, M., Gool, L.V., Williams, C.K.I., Winn, J.M., Zisserman, A.: The pascal visual object classes (VOC) challenge. IJCV **88**(2), 303–338 (2010)
7. Fang, H., Lu, G., Fang, X., Xie, J., Tai, Y., Lu, C.: Weakly and semi supervised human body part parsing via pose-guided knowledge transfer. CoRR **abs/1805.04310** (2018)
8. Gong, K., Gao, Y., Liang, X., Shen, X., Wang, M., Lin, L.: Graphonomy: Universal human parsing via graph transfer learning. In: CVPR. pp. 7450–7459 (2019)
9. Gong, K., Liang, X., Li, Y., Chen, Y., Yang, M., Lin, L.: Instance-level human parsing via part grouping network. In: ECCV. pp. 805–822 (2018)
10. Gong, K., Liang, X., Zhang, D., Shen, X., Lin, L.: Look into person: Self-supervised structure-sensitive learning and a new benchmark for human parsing. In: CVPR. pp. 6757–6765 (2017)
11. Hariharan, B., Arbeláez, P.A., Girshick, R.B., Malik, J.: Simultaneous detection and segmentation. In: ECCV. pp. 297–312 (2014)
12. He, K., Gkioxari, G., Dollár, P., Girshick, R.B.: Mask R-CNN. In: ICCV. pp. 2980–2988 (2017)
13. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR. pp. 770–778 (2016)
14. Hu, J., Shen, L., Sun, G.: Squeeze-and-excitation networks. In: CVPR. pp. 7132–7141 (2018)
15. Huang, G., Liu, Z., Weinberger, K.Q.: Densely connected convolutional networks. CoRR **abs/1608.06993** (2016)
16. Huang, Z., Wang, C., Wang, X., Liu, W., Wang, J.: Semantic image segmentation by scale-adaptive networks. TIP (2019)
17. Kimchi, R.: Primacy of wholistic processing and global/local paradigm: a critical review. Psychological bulletin **112**(1), 24 (1992)
18. Kontschieder, P., Fiterau, M., Criminisi, A., Bulò, S.R.: Deep neural decision forests. In: ICCV. pp. 1467–1475 (2015)
19. LeCun, Y., Boser, B.E., Denker, J.S., Henderson, D., Howard, R.E., Hubbard, W.E., Jackel, L.D.: Backpropagation applied to handwritten zip code recognition. Neural Computation **1**(4), 541–551 (1989)
20. Li, J., Zhao, J., Wei, Y., Lang, C., Li, Y., Feng, J.: Towards real world human parsing: Multiple-human parsing in the wild. CoRR **abs/1705.07206** (2017)

21. Liang, X., Gong, K., Shen, X., Lin, L.: Look into person: Joint body parsing & pose estimation network and A new benchmark. CoRR **abs/1804.01984** (2018)
22. Liang, X., Gong, K., Shen, X., Lin, L.: Look into person: Joint body parsing & pose estimation network and a new benchmark. TPAMI **41**(4), 871–885 (2019)
23. Liang, X., Lin, L., Shen, X., Feng, J., Yan, S., Xing, E.P.: Interpretable structure-evolving LSTM. CoRR **abs/1703.03055** (2017)
24. Liang, X., Liu, S., Shen, X., Yang, J., Liu, L., Dong, J., Lin, L., Yan, S.: Deep human parsing with active template regression. CoRR **abs/1503.02391** (2015)
25. Liang, X., Shen, X., Feng, J., Lin, L., Yan, S.: Semantic object parsing with graph LSTM. In: ECCV. pp. 125–143 (2016)
26. Lin, G., Milan, A., Shen, C., Reid, I.D.: Refinenet: Multi-path refinement networks for high-resolution semantic segmentation. In: CVPR. pp. 5168–5177 (2017)
27. Lin, T., Maire, M., Belongie, S.J., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft COCO: common objects in context. In: ECCV. pp. 740–755 (2014)
28. Liu, T., Ruan, T., Huang, Z., Wei, Y., Wei, S., Zhao, Y., Huang, T.: Devil in the details: Towards accurate single and multiple human parsing. CoRR **abs/1809.05996** (2018)
29. Long, J., Shelhamer, E., Darrell, T.: Fully convolutional networks for semantic segmentation. In: CVPR. pp. 3431–3440 (2015)
30. Luo, Y., Zheng, Z., Zheng, L., Guan, T., Yu, J., Yang, Y.: Macro-micro adversarial network for human parsing. In: ECCV (2018)
31. Nie, X., Feng, J., Yan, S.: Mutual learning to adapt for joint human parsing and pose estimation. In: ECCV. pp. 519–534 (2018)
32. Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., Wojna, Z.: Rethinking the inception architecture for computer vision. CoRR **abs/1512.00567** (2015)
33. Wang, W., Zhang, Z., Qi, S., Shen, J., Pang, Y., Shao, L.: Learning compositional neural information fusion for human parsing. In: ICCV (2019)
34. Wang, W., Zhu, H., Dai, J., Pang, Y., Shen, J., Shao, L.: Hierarchical human parsing with typed part-relation reasoning. In: CVPR (2020)
35. Wang, Y., Tran, D., Liao, Z.: Learning hierarchical poselets for human parsing. In: CVPR. pp. 1705–1712 (2011)
36. Xia, F., Wang, P., Chen, L., Yuille, A.L.: Zoom better to see clearer: Human and object parsing with hierarchical auto-zoom net. In: ECCV. pp. 648–663 (2016)
37. Xia, F., Wang, P., Chen, X., Yuille, A.L.: Joint multi-person pose estimation and semantic part segmentation. In: CVPR. pp. 6080–6089 (2017)
38. Xiao, H.: NDT: neual decision tree towards fully functioned neural graph. CoRR **abs/1712.05934** (2017)
39. Zhang, R., Tang, S., Zhang, Y., Li, J., Yan, S.: Scale-adaptive convolutions for scene parsing. In: ICCV. pp. 2050–2058 (2017)
40. Zhao, H., Shi, J., Qi, X., Wang, X., Jia, J.: Pyramid scene parsing network. In: CVPR. pp. 6230–6239 (2017)
41. Zhao, J., Li, J., Cheng, Y., Sim, T., Yan, S., Feng, J.: Understanding humans in crowded scenes: Deep nested adversarial learning and A new benchmark for multi-human parsing. In: ACM MM. pp. 792–800 (2018)
42. Zhao, J., Li, J., Cheng, Y., Zhou, L., Sim, T., Yan, S., Feng, J.: Understanding humans in crowded scenes: Deep nested adversarial learning and A new benchmark for multi-human parsing. CoRR **abs/1804.03287** (2018)
43. Zhu, X., Hu, H., Lin, S., Dai, J.: Deformable convnets V2: more deformable, better results. In: CVPR. pp. 9308–9316 (2019)