Self-Supervised Neural Aggregation Networks for Human Parsing

Jian Zhao[1,2] Jianshu Li[1] Xuecheng Nie[1] Fang Zhao[1] Yunpeng Chen [1] Zhecan Wang [3] Jiashi Feng [1] Shuicheng Yan[1,4]
[1] National University of Singapore [2] National University of Defense Technology [3] Franklin. W. Olin College of Engineering [4] 360 AI Institute
{zhaojian90, jianshu, niexuecheng, chenyunpeng}@u.nus.edu zhecan.wang@students.olin.edu {elezhf, elefjia, eleyans}@nus.edu.sg

Abstract

In this paper, we present a Self-Supervised Neural Aggregation Network (SS-NAN) for human parsing. SS-NAN adaptively learns to aggregate the multi-scale features at each pixel "address". In order to further improve the feature discriminative capacity, a self-supervised joint loss is adopted as an auxiliary learning strategy, which imposes human joint structures into parsing results without resorting to extra supervision. The proposed SS-NAN is end-to-end trainable. SS-NAN can be integrated into any advanced neural networks to help aggregate features regarding the importance at different positions and scales and incorporate rich high-level knowledge regarding human joint structures from a global perspective, which in turn improve the parsing results. Comprehensive evaluations on the recent Look into Person (LIP) and the PASCAL-Person-Part benchmark datasets demonstrate the significant superiority of our method over other state-of-the-arts.

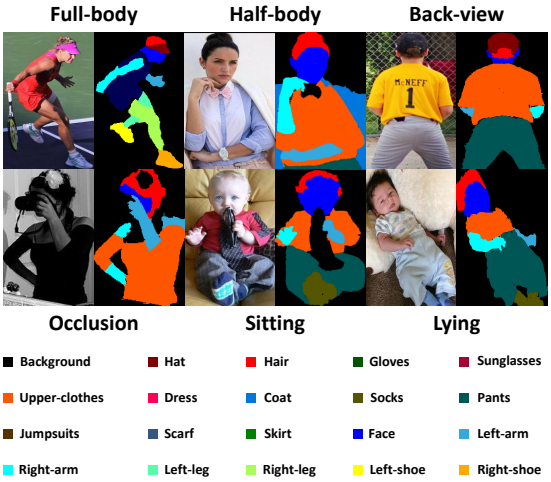

Figure 1. Examples for human parsing by our SS-NAN architecture. Best viewed in color.

1. Introduction

Human parsing, also known as human semantic segmentation, relates to the problem of assigning fine-grained semantic labels (*e.g.* "hair", "face", "dress", *etc.*) to every pixel in the image, as illustrated in Figure 1. It is a very challenging computer vision task and one of the most crucial steps towards detailed image understanding for human-centric analysis. Successful human parsing techniques could facilitate huge higher-level artificial intelligence applications, such as human behavior analysis [11, 25], clothing style recognition and retrieval [8], and automatic product recommendation [20].

Recently, deep learning methods, and in particular Fully Convolutional Networks (FCNs) [31] based methods, *e.g.*, Segnet [1], DeepLabV2 [27], Attention [4], have shown remarkable success in human parsing on several benchmarks. One of the key elements to successful human parsing among these models is the use of multi-scale features. Compared to the counterparts with single-scale, multi-scale based methods perform a human-like "auto-zoom" process to exploit diverse contextual information from global and local regions, which compensate each other and naturally benefits the human parsing task to solve. However, it is still an open problem to build an appropriate representation of the multi-scale features, such that it can effectively incorporate useful information across different scales together, maintaining beneficial while discarding noise.

One intuitive approach would be combining features from the intermediate layers of FCNs, which we refer to as Skip-Net and they have dominated human parsing recently [31, 26, 38, 30]. Hierarchical features within a Skip-Net are multi-scale in nature due to the increasing receptive field sizes. Such a combined representation comprehensively maintains the information across all scales. However, to train a Skip-Net, one usually needs to employ a two-stage process [31, 26, 38, 30] by first training the backbone model with a classifier and then slightly fine-tuning during multi-scale feature extraction. Such training process is not ideal due to separate stages and expensive time cost (*e.g.*, three to six days [26]).

We argue that it is more desirable to come with a com-

pact and static feature representation across different scales, irrespective of the varied number of scales. Such representation learning would allow a more efficient training process. A straightforward solution might be Share-Net, which resizes the input image to several scales and passes each through a shared FCN. The resulting multi-scale features are then aggregated together by conducting a certain type of pooling to form the final dense prediction[12, 28].

The most commonly adopted pooling strategies may be average- or max-pooling over scales [10, 21, 2]. While these intuitive pooling strategies were shown to be effective in the previous works, features at each scale are either treated equally or selected sparsely. We believe that a good aggregation strategy should adaptively weigh and aggregate the features across all scales. The intuition is simple: a human image may contain several big and small semantic fashion / body regions (*e.g.*, upper-clothes *v.s.* sunglasses), and a smart algorithm should favor features that are more relevant (or more "needed") and prevent noise from jeopardizing human parsing.

To this end, we look for an adaptive weighting scheme to spatially aggregate all features from several different scales together to form a compact and static representation. Different from previous methods, we neither fix the weights for addressing feature map values nor rely on any particular heuristics to set them. Instead, we design a neural network to adaptively learn the weights.

Our neural aggregation network is designed to inherit the main advantages of pooling techniques, including the ability to handle arbitrary scale number and producing order-invariant representations. The idea is inspired by the Neural Turing Machines [15], which applied an attention mechanism to organize the input through accessing an external memory. This mechanism can take arbitrary number of input and work as a tailor emphasizing or suppressing each input element via a weighted averaging, and very importantly it is order independent and has trainable parameters. In this work, we employ a Share-Net associated with this adaptive weighting mechanism for multi-scale feature aggregation.

Motivated by [14], in order to explicitly enforce the produced parsing results to be semantically consistent with the human joint structures, in addition to using the conventional pixel-wise part annotations as the supervision, we further employ a joint loss to enhance the quality of predicted parsing results from a joint structure perspective. That means a satisfactory parsing result should be able to preserve a reasonable spatial joint layout structure of human parts. We generate approximated human joints directly from the parsing annotations and use them as the auxiliary supervision signal for the joint loss, which is hence a "self-supervised" strategy. We thus term our approach as the Self-Supervised Neural Aggregation Network (SS-NAN), which can be trained in an end-to-end way for human parsing. The SS-NAN not only enjoys the time and memory efficiency due to the adaptively aggregated representations, but also exhibits superior performance, as we will show in our experiments.

We demonstrated the effectiveness of the proposed SS-NAN on the challenging human parsing benchmarks, including Look into Person (LIP) [14] and PASCAL-Person-Part [5]. Experimental results show that the proposed SS-NAN consistently improves over strong baselines. More importantly, the proposed SS-NAN can serve as a general framework for learning multi-scale adaptive pooling. Therefore, it may also serve as a feature aggregation scheme for other computer vision tasks.

Our contributions can be summarized in the following three aspects:

- We propose an effective Self-Supervised Neural Aggregation Network (SS-NAN) for human parsing, which adaptively aggregates the multi-scale features while explicitly enforcing consistency between the parsing results and the human joint structures. Our model is flexible in that it can be modified in various ways.
- The proposed SS-NAN can be effectively trained in an end-to-end way without any separate pre-processing or post-processing, which is crucial for best human parsing performance.
- The proposed SS-NAN significantly surpasses the previous methods on both challenging LIP [14] and PASCAL-Person-Part benchmark datasets [5].

2. Related Works

Our model draws success from several areas, including FCNs, multi-scale features for human parsing, and feature aggregation schemes.

FCNs: FCNs [31] based methods have demonstrated state-of-the-art performance for the human parsing problem, including [31, 26, 38, 30, 4]. Our method works directly on the pixel-level representation, similar to many some recent research on semantic image segmentation. Farabet *et al.* [9] proposed a multi-scale FCN based framework appended with a dense pixel-level Conditional Random Field (CRF) for pixel-wise labeling. Chen *et al.* [4] proposed an attention mechanism that learns to weight the multi-scale features at each pixel location. The main difference between our proposed SS-NAN and these previous methods is the seamless integration of the state-of-the-art backbone Skip-Net (*e.g.*, ResNet-101 [17]) and Share-Net architecture, the adaptive neural aggregation scheme across multi-scale features, and the auxiliary self-supervised structure-sensitive learning strategy in consistence with human joint structures into an end-to-end trainable unified network, which is very important in boosting

the human parsing performance as demonstrated in the experiments.

Multi-scale features for human parsing: Multi-scale features have been shown significantly useful for computer vision problems and in particular for human parsing, which enables the network implicitly "look into" the most important information within an image. There are several existing methods which exploit multi-scale features for semantic image segmentation. Skip-Net exploits hierarchical multi-scale features from different levels of the network. Hypercolumn [16] merges multi-scale features from intermediate layers and learns the final dense prediction through a stage-wise training process instead of end-to-end training. SegNet [1] and U-Net [33] apply skip-connections in the deconvolution architecture to exploit the multi-scale features from intermediate layers. Share-Net exploits multi-scale features by applying multi-scale input images to a shared network. Farabet *et al.* [9] employed a Laplacian pyramid, passed each scale through a shared network, and fused the features from all the scales. Lin *et al.* [28] resized the input images into three scales and concatenated the resulting three-scale features to generate the unary and pair-wise potentials of a CRF. Although there are many existing work exploiting multi-scale features for human parsing, few of them provide satisfactory results.

Feature aggregation schemes: For multi-scale feature aggregation, existing methods either use average-pooling [6, 7] or max-pooling [10, 32] over different scales. Motivated by [14], we propose to jointly learn an adaptive neural aggregation model that softly weights the features from different input scales when predicting the semantic label of a pixel. The final dense prediction is produced by the aggregated probability maps across all the scales. Incorporated with a self-supervised structure-sensitive learning approach, the proposed SS-NAN leverages human joint structure more effectively and efficiently, which can be modified in various ways. As previously mentioned, this work is also related to the Neural Turing Machines [15]. However, it is worth noting that we only borrow their differentiable memory addressing scheme for our multi-scale feature aggregation.

3. Self-Supervised Neural Aggregation Network

As shown in Figure 2, our proposed SS-NAN seamlessly integrate the state-of-the-art backbone Skip-Net and Share-Net architecture for multi-scale feature learning, the adaptive neural aggregation scheme across various scales, the pixel-wise softmax loss and an auxiliary self-supervised joint loss in consistence with human joint structures into an end-to-end trainable unified network, which is beneficial for boosting the human parsing performance. We now present each component in detail.

3.1. Multi-scale feature learning

As noted previously, we aim to exploit multi-scale features to "look into" the most important information for human parsing. SS-NAN provides a generic means to learn hierarchical multi-scale features. A crucial aspect of the design ensures that the gradient can be effortlessly backpropagated through the network all the way to low-level layers over multiple skip connections with shared weights between each scale embedding, ensuring that the entire network can be trained end-to-end.

Deeper networks have shown to yield better performance for many computer vision problems [17, 21, 35]. However, naively increasing depth of the network may introduce additional optimization difficulty as stated in [34, 23]. An effective solution to this problem is Skip-Net (*e.g.*, ResNet-101 [17]), which adds a skip-connection that bypasses the non-linear transformations with an identity function:

$$x_l = f_l(x_{l-1}) + x_{l-1}, \tag{1}$$

where x_l is the output of the l^{th} layer, $f_l(\cdot)$ can be a composite non-linear transformations, such as Convolution (Conv), Batch Normalization (BN) [18], Rectified Linear Units (ReLU) [13], or Pooling [22].

As FCNs [31] based methods have proven significantly successful in human parsing [31, 26, 38, 30], we further modify the Skip-Net architecture to be fully convolutional, which captures the cross-layer context, facilitates the gradient Backpropagation (BP), and produces reasonable dense predictions. In particular, the last fully-connected layers are turned into convolutional layers (*e.g.*, the last layer has a spatial convolutional kernel with size $C \times 3 \times 3$, where C is the number of human parsing classes of interest).

As shown in Figure 2, our basic multi-scale learning archiecture is a Share-Net, which incorporates three FCN based Skip-Nets with shared weights. It simultaneously considers the local fine details and global structure information. The input to our proposed SS-NAN is a 321×321 color image and then resized to three different scales / resolutions $s \in \{0.5, 0.75, 1.0\}$ according to [14]. The resulted multi-scale input are then fed to the subsequent parallel FCN based Skip-Nets for multi-scale feature learning. More details are provided in Sec. 4.

3.2. Adaptive neural aggregation

Herein, we discuss how to adaptively aggregate the multi-scale features for our proposed model.

Based on the above-mentioned Share-Net, each scale input is passed through the corresponding Skip-Net (the FCN weights are shared across all scales), and produces a probability map for scale s, denoted as $x^s_{i,c}$ where i ranges over all the spatial pixel-wise "address" and $c \in \{1, \cdots, C\}$. The probability map $x^s_{i,c}$ is then resized to the same resolution *w.r.t.* the finest scale by bilinear interpolation. Our goal is

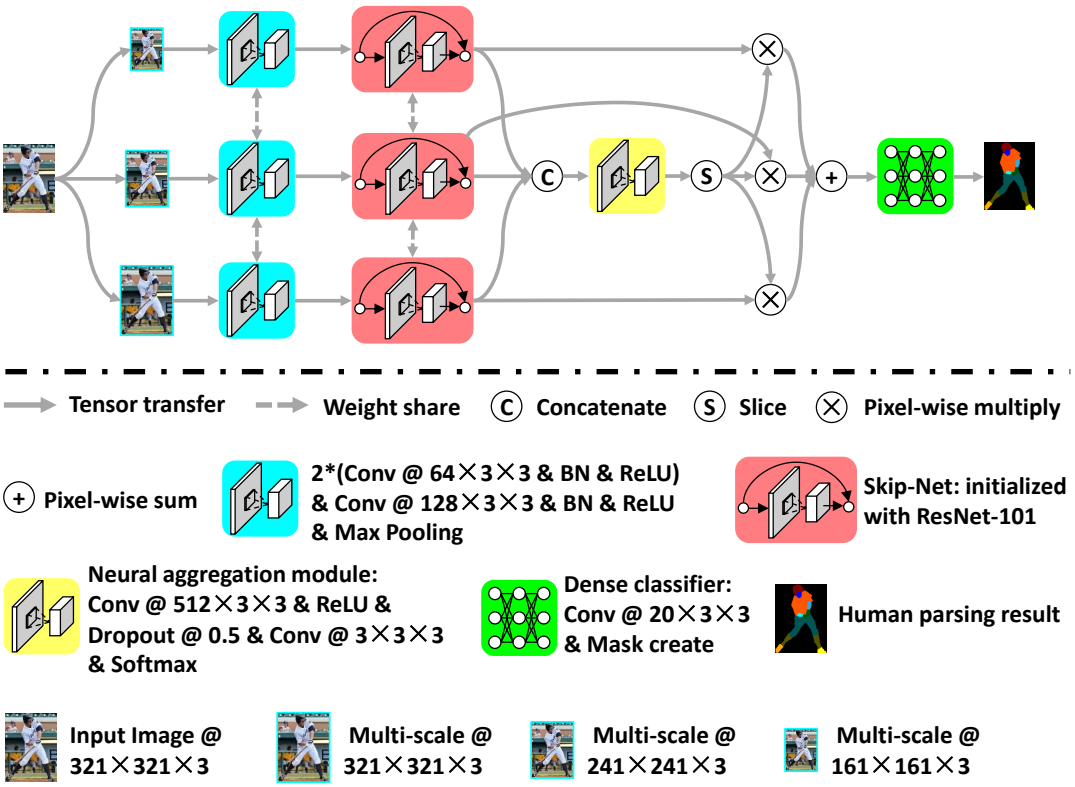

Figure 2. Overview of the proposed SS-NAN architecture. SS-NAN takes the human images as input and outputs the corresponding dense parsing results. It seamlessly integrate the state-of-the-art backbone Skip-Net and Share-Net architecture and the adaptive neural aggregation scheme across various scales for multi-scale feature learning and adaptive neural aggregation, which is significantly beneficial for the human parsing problem. Best viewed in color.

to utilize all probability maps from all scales to generate a compact and static feature representation, *i.e.*,

$$r_{i,c} = \sum_{s}^{S} w_i^s \cdot x_{i,c}^s, \tag{2}$$

where r denotes the aggregated probability map at (i, c) for all scales.

Obviously, the key of Eq. 2 is its weights $\{w_i^s\}$. If $w_i^s \equiv \frac{1}{N_S}$, Eq. 2 will degrades to naive average-pooling, which is usually non-optimal as the naive max-pooling. We instead seek to employ a better aggregation scheme.

Two main principles have been considered in our adaptive neural aggregation module. First, the module should be able to process different number of scales and invariant to the scale order for further modification and generic use. Second, the module should be adaptive to the learned multi-scale features and has parameters end-to-end trainable through the standard Stochastic Gradient Descent (SGD) and BP algorithm.

Our solution is inspired by the memory addressing mechanism described in [15]. The idea therein is to use a neural model to read external memories through a differentiable addressing scheme, which is applicable to our adaptive aggregation scenario. In this work, we treat the multi-scale features as the memory and cast adaptive weighting as a memory addressing procedure. Our neural aggregation module reads all scale feature tensors from the multi-scale feature learning module, and adaptively generate linear weights for them for aggregation. In particular, the weight w_i^s is computed by

$$w_i^s = \frac{exp(a_i^s)}{\sum_s^S exp(a_i^s)}, \tag{3}$$

where a_i^s is the feature map produced by the adaptive neural aggregation module at (i, s). Note that w_i^s is shared across all channels. The adaptive neural aggregation module takes as input the multi-scale probability maps from each Skip-Net, and it consists of two Conv layers with the kernel size $512 \times 3 \times 3$ and $N_S \times 3 \times 3$, repectively.

It can be seen that our adaptive neural aggregation algorithm essentially selects one point inside of the convex hull spanned by all the multi-scale features. In this way, the number and the order of scales do not affect the aggregation

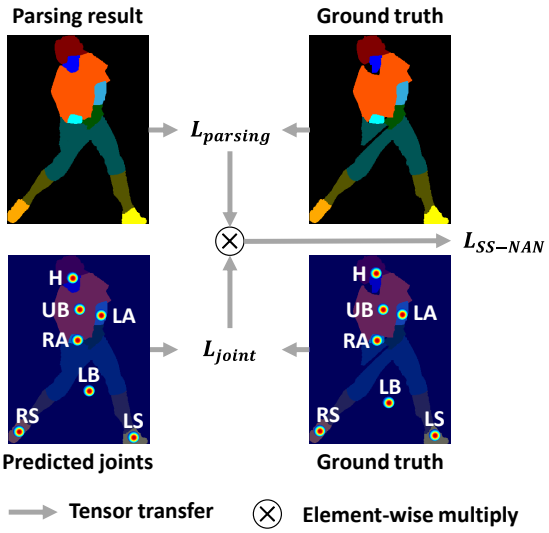

Figure 3. Illustration of our optimization strategy for the proposed SS-NAN. The generated joints and joints ground truth coordinates are obtained by computing the center points of corresponding regions in parsing maps, including head (H), upper-body (UB), lower-body (LB), right-arm (RA), left-arm (LA), right-leg (RL), left-leg (LL), right-shoe (RS), left-shoe (LS). The final loss function for SS-NAN is generated by weighting the pixel-wise Cross-Entropy loss with the joint loss. Best viewed in color.

results. The weight w_i^s reflects the importance of feature at (i, s). As a result, the neural aggregation module adaptively descides how much attention to pay to features at different "addresses" and scales. We emphasize that the adaptive neural aggregation module computes a soft weight for each "address" and scale, and it allows the gradient to be back-propagated through, similar to [15]. One significant advantage is that tedious annotation of the "ground truth scale" for each pixel is avoided, allowing the neural aggregation module adaptively learn the best weights on scales.

3.3. Optimization

We optimize the network parameters using training images with pixel-wise annotations. The final output is produced by performing a softmax operation on the aggregated probability maps across all scales. Inspired by [14], in order to enforce the human parsing results to be semantically consistent with the human joint structures, in addition to the conventional pixel-wise Cross-Entropy loss, we further employ a structure-sensitive learning strategy, which is a self-supervised strategy without any expensive human pose annotation. Nine joints are defined to construct the human joint structure, *i.e.*, the centers of regions of head, upper body, lower body, left arm, right arm, left leg, right leg, left shoe, and right shoe, as shown in Figure 3. Each region is generated by merging parsing labels of several related small regions (*e.g.*, the region of head is merged by regions of hat, hair, sunglasses, and face). For each parsing result and the corresponding ground truth, the centers of regions are computed dynamically to obtain joint coordinates. Then an Euclidean (L2) distance is used to measure the quality of the generated joint structures, which also reflect the human joint layout consistency between the predicted parsing results and the ground truth. Finally, we minimize the loss function weighted by the pixel-wise Cross-Entropy loss averaged over all pixel "addessses" and the joint loss with the standard SGD and BP algorithm. More formally, the final loss function for our proposed SS-NAN is calculated as:

$$\mathcal{L}_{SS-NAN} = \mathcal{L}_{parsing} \cdot \mathcal{L}_{joint}, \tag{4}$$

where $\mathcal{L}_{parsing}$ is the pixel-wise Cross-Entropy loss calculated based on the parsing annotations, $\mathcal{L}_{joint} = \frac{1}{2N_{joint}} \sum_{i=1}^{N_{joint}} \|J_i^p - J_i^{gt}\|_2^2$, J_i^p is the i^{th} joint coordinates computed according to the dense predictions, J_i^{gt} is the joint coordinates obtained from corresponding parsing ground truth, N_{joint} is the pre-defined joint number 9.

In addition to the supervision to the final output, extra supervision (*i.e.*, dense classifier Conv $20 \times 3 \times 3$ & $\mathcal{L}_{SS-NAN}$) is injected to the final output of each Skip-Net (*i.e.*, each scale) within the Share-Net. Such deeply supervised strategy allows the proposed SS-NAN to be trained effectively and efficiently in an end-to-end way.

4. Experiments

4.1. Experimental settings

Benchmark datasets: We evaluate the performance of our proposed SS-NAN for human parsing on the challenging LIP [14] and PASCAL-Person-Part [5] public benchmark datasets.

LIP benchmark dataset[1] **[14].** To further push the frontiers of semantic image segmentation and in particular human parsing research, recently Liang *et al.* [14] developed and publicly released a new large-scale benchmark dataset Look into Person (LIP) focusing on semantic fine-grained understanding of human bodies, which makes a significant advance in terms of scalability, diversity and difficulty. LIP is an order of magnitude larger and more challenging than previous similar attempts [5, 26, 37]. The images in the LIP dataset are cropped person instances from Microsoft COCO [29] training and validation sets. Thus, the images of LIP are collected from the real-world scenarios containing people appearing with challenging poses, viewpoints, heavy occlusions, various appearences and in wide range of resolutions. Moreover, the background of images of LIP is also more complex and divese than the one in previous counterparts. LIP is well-annotated with elaborated pixel-wise

[1] http://hcp.sysu.edu.cn/lip/.

annotations with 19 semantic human part labels (*i.e.*, hat, hair, gloves, sunglasses, upper-clothes, dress, coat, socks, pants, jumpsuits, scarf, skirt, face, left-arm, right-arm, left-leg, right-leg, left-shoe, and right-shoe) and one background label. There are $50,462$ images in the LIP dataset, including $19,081$ full-body images, $13,672$ upper-body images, 403 lower-body images, $3,386$ head-missed images, $2,778$ back-view images, and $21,028$ images with occlusions. LIP is further split into separate training set containing $30,462$ images, validation set containing $10,000$ images, and testing set containing $10,000$ images. The annotations for testing set is officially withheld for benchmarking purpose.

PASCAL-Person-Part benchmark dataset[2] **[5].** PASCAL-Person-Part benchmark dataset is a set of additional annotations for PASCAL VOC 2010. It goes beyond the original PASCAL object detection task by providing fine-grained pixel-wise labels for six body part of the human, *i.e.*, head, torso, upper- / lower-arms, and upper- / lower-legs. The rest of each image is considered as background. There are $3,535$ images in the PASCAL-Person-Part dataset, which is split into separate training set containing $1,717$ images and testing set containing $1,818$ images.

Metrics: We report three metrics for human parsing that are variations on pixel accuracy and region Intersection over Union (IoU).

- Pixel accuracy: $\frac{\sum_i n_{ii}}{\sum_i t_i}$,
- Mean accuracy: $\frac{1}{C}\sum_i \frac{n_{ii}}{t_i}$,
- Mean IoU: $\frac{1}{C}\sum_i \frac{n_{ii}}{t_i+\sum_j n_{ji}-n_{ii}}$,

where n_{ji} is the number of pixels of class i predicted to belong to class j, C is the parsing class number, $t_i = \sum_j n_{ij}$ is the total number of pixels of class i.

Network architecture: The input to our proposed SS-NAN is a 321×321 color image and then resized to three different scales $s \in \{0.5, 0.75, 1.0\}$. Our Skip-Net backbone architecture for each scale within the Share-Net is initialized from the publicly available model, Pyramid Scene Parsing (PSP) network [39] due to its leading accuracy and competitive efficiency, with slight modification by dropping the top pyramid pooling module and dense prediction module. Thus, each Skip-Net in our SS-NAN becomes a ResNet-101[17] with PSP pre-trained weights. The adaptive aggregation module in our SS-NAN consists of two Conv layers with the kernel size $512 \times 3 \times 3$ and $N_S \times 3 \times 3$, repectively.

Training: SDG with mini-batch is used for training. We set the mini-batch size of 30 images. Inspired by [3], we use the "poly" learning rate policy where the current learning rate equals to the base one multiplying $(1 - \frac{iter}{max\_iter})^{power}$. We set the base learning rate of 0.001 (0.01 for the final dense classifier layer) and $power$ to 0.9. We use the momentum of 0.9 and weight decay of 0.0005. For data augmentation, we adopt random mirror and random resize between 0.6 and 1.4 for all datasets. This comprehensive data augmentation scheme makes the network resist overfitting. Following [14], two training steps are employed to optimize our SS-NAN. First, we train the basic network with only $\mathcal{L}_{parsing}$ for 40 epoches, which takes about three and a half days. Then we perform "self-supervised" strategy to fine-tune our model with the $\mathcal{L}_{SS-NAN}$ for roughly 30 epoches and it takes about two days. In the testing stage, one image takes 0.5 second on average.

[2] http://www.stat.ucla.edu/ xianjie.chen/pascal_part_dataset/pascal_part.html.

Method	Overall accuracy	Mean accuracy	Mean IoU
SegNet [1]	69.04	24.00	18.17
FCN-8s [31]	76.06	36.75	28.29
DeepLabV2 [27]	82.66	51.64	41.64
Attention [4]	83.43	54.39	42.92
Attention+SSL [14]	84.36	54.94	44.73
SS-NAN (ours)	**87.59**	**56.03**	**47.92**

Table 1. Performance comparison of SS-NAN with five state-of-the-art methods on the LIP validation set. The best performance is highlighted in bold.

Reproducibility: The proposed method is implementated by extending the Caffe framework [19]. All networks are trained on a single NVIDIA GeForce GTX TITAN X GPU with 12GB memory. The source code and trained models for our SS-NAN will be released.

4.2. Results and comparisons

We compare the proposed method with the strong baselines on the two challenging public benchmark datasets.

LIP benchmark dataset [14]. We report the quantitive results and comparisons with five state-of-the-art methods on LIP validation set in Table 1. We can observe that the proposed SS-NAN achieves a huge boost in average IoU: 3.19% better than Attention+SSL [14] and 5.00% better than Attention [4]. This superior performance of SS-NAN demonstrates the effectiveness of our multi-scale feature learning (*i.e.*, automatically "looks into" the most important information for human parsing), adaptive neural aggregation (*i.e.*, adaptively descides how much attention to pay to features at different "addresses" and scales), and self-supervised structure-sensitive learning (*i.e.*, incorporates the human joint structure into the pixel-wise dense prediction).

We further report per-class IoU on LIP validation set to verify the detailed effectiveness of our SS-NAN, presented in Table 2. With the carefully designed multi-scale feature learning module, adaptive neural aggregation module, and

Method	Hat	Hair	Gloves	Sunglasses	Upper-clothes	Dress	Coat	Socks	Pants	Jumpsuits	Scarf	Skirt	Face	Left-arm	Right-arm	Left-leg	Right-leg	Left-shoe	Right-shoe	Background	Mean IoU
SegNet [1]	26.60	44.01	0.01	0.00	34.46	0.00	15.97	3.59	33.56	0.01	0.00	0.00	52.38	15.30	24.23	13.82	13.17	9.26	6.47	70.62	18.17
FCN-8s [31]	39.79	58.96	5.32	3.08	49.08	12.36	26.82	15.66	49.41	6.48	0.00	2.16	62.65	29.78	36.63	28.12	26.05	17.76	17.70	78.02	28.29
DeepLabV2 [27]	57.94	66.11	28.50	18.40	60.94	23.17	47.03	34.51	64.00	22.38	14.29	18.74	69.70	49.44	51.66	37.49	34.60	28.22	22.41	83.25	41.64
Attention [4]	58.87	66.78	23.32	19.48	63.20	29.63	49.70	35.23	66.04	24.73	12.84	20.41	70.58	50.17	54.03	38.35	37.70	26.20	27.09	84.00	42.92
Attention+SSL [14]	59.75	67.25	28.95	21.57	65.30	29.49	51.92	38.52	68.02	24.48	14.92	24.32	71.01	52.64	55.79	40.23	38.80	28.08	29.03	84.56	44.73
SS-NAN (Ours)	**63.86**	**70.12**	**30.63**	**23.92**	**70.27**	**33.51**	**56.75**	**40.18**	**72.19**	**27.68**	**16.98**	**26.41**	**75.33**	**55.24**	**58.93**	**44.01**	**41.87**	**29.15**	**32.64**	**88.67**	**47.92**

Table 2. Performance comparison *w.r.t.* per-class mean IoU of SS-NAN with five state-of-the-art methods on the LIP validation set. The best performance is highlighted in bold.

Method	Head	Torso	Upper-arms	Lower-arms	Upper-legs	Lower-legs	Background	Mean IoU
DeepLabV2 [27]	78.09	54.02	37.29	36.85	33.73	29.61	92.85	51.78
HAZN [36]	80.79	59.11	43.05	42.76	38.99	34.46	93.59	56.11
Attention [4]	81.47	59.06	44.15	42.50	38.28	35.62	93.65	56.39
LG-LSTM [24]	82.72	60.99	45.40	**47.76**	42.33	37.96	88.63	57.97
Attention+SSL [14]	83.26	62.40	47.80	45.58	42.32	39.48	94.68	59.36
SS-NAN (ours)	**86.43**	**67.28**	**51.09**	**48.07**	**44.82**	**42.15**	**97.23**	**62.44**

Table 3. Performance comparison *w.r.t.* per-class mean IoU of SS-NAN with five state-of-the-art methods on the PASCAL-Person-Part benchmark dataset. The best performance is highlighted in bold.

optimization scheme, we achieved the best performance on all the classes. As observed from the reported results, SS-NAN significantly improves the performance of the labels like arms, legs, and shoes, which demonstrates its excellent ability to distinguish "left" *v.s.* "right". Furthermore, the labels covering small regions such as sunglasses, gloves, socks, are predicted better with higher IoU. This improvement also verified the effectiveness of the proposed SS-NAN especially for small labels.

PASCAL-Person-Part benchmark dataset [5]. Table 3 shows the performance of our SS-NAN and comparisons with five state-of-the-art methods on the standard mean IoU criterion. Our method can significantly outperform all baselines. In particular, our SS-NAN achieves mean IoU of 62.44%, 3.08% better than Attention+SSL [14] and 4.47% better than LG-LSTM [24]. This huge improvement demonstrates that our proposed SS-NAN is significantly beneficial for human parsing with the combination of Skip-Net, Share-Net, multi-scale feature learning, adaptive neural aggregation, and deep supervision.

Note that without self-supervision via human joint prediction, around 1% performance decrease can be observed on both benchmarks, which proves its effectiveness and benefits for human parsing.

4.3. Qualitative comparison

The qualitative comparisons of human parsing results on the LIP validation set are visualized in Figure 4. As can be observed, our SS-NAN outputs more semantically reasonable, meaningful and precise predictions than Attention+SSL [14] despite the existence of large pose, view-

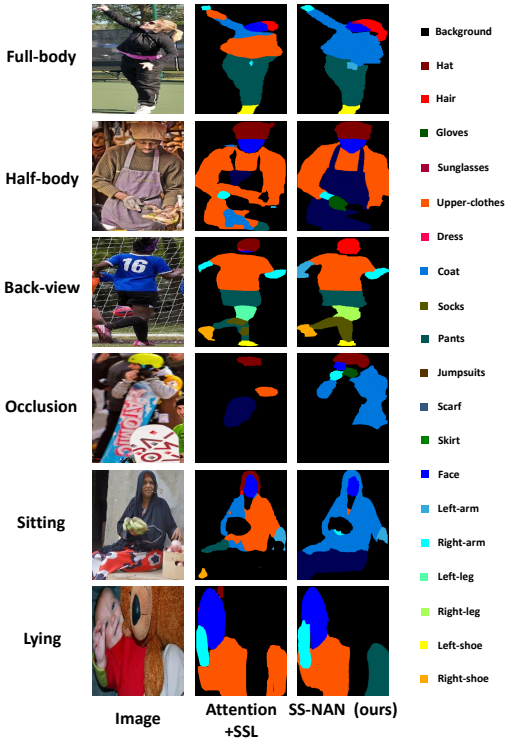

Figure 4. Visualized comparison of human parsing results on the LIP validation set. Best viewed in color.

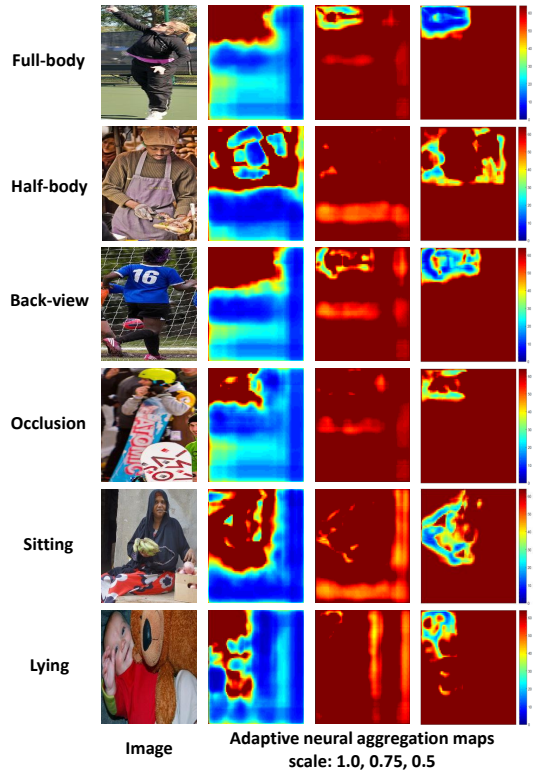

Figure 5. Visualized adaptive neural aggregation maps *w.r.t.* three different scales $s \in \{0.5, 0.75, 1.0\}$ learned by our SS-NAN. Best viewed in color.

point, occlusion, appearance and resolution variations. For example, observed from the full-body image, the small regions (*e.g.*, sunglasses and left-shoe) can be successfully segmented out by our method. Taking half-body and sitting images for example, our approach can also successfully handle the confusing labels such as upper-clothes, coat, and scarf. These regions with similar appearances can be recognized and separated by the guidance from local and global multi-scale information and human joint structure information. For the most difficult back-view and occlusion images, the left-arm, right-arm, gloves, left-shoe, right-shoe, and part of the right-leg are excellently predicted and masked out by our approach. In general, by effectively exploiting multi-scale information and human joint structure information with the carefully designed network architecture, our approach outputs more accurate results for confusing labels on the human parsing task.

Moreover, in order to gain insight into the adaptive neural aggregation mechanism, we further visualized the pixel-wise weight maps *w.r.t.* three different scales $s \in \{0.5, 0.75, 1.0\}$ learned by our SS-NAN in Figure 5. As can be observed, our SS-NAN has adaptively learned to put higher weights on small regions for scale 1.0, on middle regions for scale 0.75, and on large regions and background for scale 0.5. Such a human-like "auto-zoom" process exploits diverse contextual information from global and local regions, which compensate each other and naturally benefits the human parsing task to solve.

5. Conclusion

In this paper, we proposed an effective and efficient Self-Supervised Neural Aggregation Network (SS-NAN) for human parsing. SS-NAN learns the comprehensive multi-scale features through a Share-Net containing Skip-Nets for each scale stream. The multi-scale features are adaptively aggregated while explictly enforcing consistency between the parsing results and the human joint structures. SS-NAN can be effortlessly optimized in an end-to-end way with the pixel-wise Cross-Entropy loss and an auxiliary joint loss and can be generalized to more real-world applications. Extensive evaluations on the two challenging human parsing benchmark datasets (*i.e.*, LIP and PASCAL-Person-Part) clearly verified the effectiveness of the proposed approach.

Acknowledgement

The work of Jian Zhao was partially supported by China Scholarship Council (CSC) grant 201503170248.

The work of Jiashi Feng was partially supported by National University of Singapore startup grant R-263-000-C08-133 and Ministry of Education of Singapore AcRF Tier One grant R-263-000-C21-112.

References

[1] V. Badrinarayanan, A. Kendall, and R. Cipolla. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *arXiv preprint arXiv:1511.00561*, 2015.

[2] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille. Semantic image segmentation with deep convolutional nets and fully connected crfs. *arXiv preprint arXiv:1412.7062*, 2014.

[3] 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. *arXiv preprint arXiv:1606.00915*, 2016.

[4] L.-C. Chen, Y. Yang, J. Wang, W. Xu, and A. L. Yuille. Attention to scale: Scale-aware semantic image segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3640–3649, 2016.

[5] 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 *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1971–1978, 2014.

[6] D. Ciregan, U. Meier, and J. Schmidhuber. Multi-column deep neural networks for image classification. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 3642–3649. IEEE, 2012.

[7] J. Dai, K. He, and J. Sun. Boxsup: Exploiting bounding boxes to supervise convolutional networks for semantic segmentation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1635–1643, 2015.

[8] W. Di, C. Wah, A. Bhardwaj, R. Piramuthu, and N. Sundaresan. Style finder: Fine-grained clothing style detection and retrieval. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 8–13, 2013.

[9] C. Farabet, C. Couprie, L. Najman, and Y. LeCun. Learning hierarchical features for scene labeling. *IEEE transactions on pattern analysis and machine intelligence*, 35(8):1915–1929, 2013.

[10] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part-based models. *IEEE transactions on pattern analysis and machine intelligence*, 32(9):1627–1645, 2010.

[11] C. Gan, M. Lin, Y. Yang, G. de Melo, and A. G. Hauptmann. Concepts not alone: Exploring pairwise relationships for zero-shot video activity recognition. 2016.

[12] R. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 580–587, 2014.

[13] X. Glorot, A. Bordes, and Y. Bengio. Deep sparse rectifier neural networks.

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

[15] A. Graves, G. Wayne, and I. Danihelka. Neural turing machines. *arXiv preprint arXiv:1410.5401*, 2014.

[16] B. Hariharan, P. Arbeláez, R. Girshick, and J. Malik. Hypercolumns for object segmentation and fine-grained localization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 447–456, 2015.

[17] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. *arXiv preprint arXiv:1512.03385*, 2015.

[18] S. Ioffe and C. Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv preprint arXiv:1502.03167*, 2015.

[19] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell. Caffe: Convolutional architecture for fast feature embedding. *arXiv preprint arXiv:1408.5093*, 2014.

[20] Y. Kalantidis, L. Kennedy, and L.-J. Li. Getting the look: clothing recognition and segmentation for automatic product suggestions in everyday photos. In *Proceedings of the 3rd ACM conference on International conference on multimedia retrieval*, pages 105–112. ACM, 2013.

[21] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.

[22] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

[23] C.-Y. Lee, S. Xie, P. Gallagher, Z. Zhang, and Z. Tu. Deeply-supervised nets. In *Artificial Intelligence and Statistics*, pages 562–570, 2015.

[24] X. Liang, X. Shen, D. Xiang, J. Feng, L. Lin, and S. Yan. Semantic object parsing with local-global long short-term memory. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3185–3193, 2016.

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

[26] 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 *Proceedings of the IEEE International Conference on Computer Vision*, pages 1386–1394, 2015.

[27] C. Liang-Chieh, G. Papandreou, I. Kokkinos, K. Murphy, and A. Yuille. Semantic image segmentation with deep convolutional nets and fully connected crfs. In *International Conference on Learning Representations*, 2015.

[28] G. Lin, C. Shen, A. van den Hengel, and I. Reid. Efficient piecewise training of deep structured models for semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3194–3203, 2016.

[29] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft coco: Common objects in context. In *European Conference on Computer Vision*, pages 740–755. Springer, 2014.

[30] 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 *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1419–1427, 2015.

[31] J. Long, E. Shelhamer, and T. Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3431–3440, 2015.

[32] G. Papandreou, I. Kokkinos, and P.-A. Savalle. Untangling local and global deformations in deep convolutional networks for image classification and sliding window detection. *arXiv preprint arXiv:1412.0296*, 2014.

[33] O. Ronneberger, P. Fischer, and T. Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 234–241. Springer, 2015.

[34] L. Shen, Z. Lin, and Q. Huang. Relay backpropagation for effective learning of deep convolutional neural networks. In *European Conference on Computer Vision*, pages 467–482. Springer, 2016.

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

[36] 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 *European Conference on Computer Vision*, pages 648–663. Springer, 2016.

[37] K. Yamaguchi, M. H. Kiapour, L. E. Ortiz, and T. L. Berg. Parsing clothing in fashion photographs. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 3570–3577. IEEE, 2012.

[38] L. Yang, H. Rodriguez, M. Crucianu, and M. Ferecatu. Fully convolutional network with superpixel parsing for fashion web image segmentation. In *International Conference on Multimedia Modeling*, pages 139–151. Springer, 2017.

[39] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia. Pyramid scene parsing network. *arXiv preprint arXiv:1612.01105*, 2016.