Hindawi
Mathematical Problems in Engineering
Volume 2020, Article ID 5794283, 12 pages
https://doi.org/10.1155/2020/5794283

Research Article

Multiscale Meets Spatial Awareness: An Efficient Attention Guidance Network for Human Parsing

Fan Zhou, Enbo Huang, Zhuo Su, and Ruomei Wang

School of Data and Computer Science, National Engineering Research Center of Digital Life, Sun Yat-sen University, Guangzhou, China

Correspondence should be addressed to Zhuo Su; suzhuo3@mail.sysu.edu.cn

Received 19 May 2020; Revised 23 September 2020; Accepted 1 October 2020; Published 17 October 2020

Academic Editor: Maria Patrizia Pera

Human parsing, which aims at resolving human body and clothes into semantic part regions from an human image, is a fundamental task in human-centric analysis. Recently, the approaches for human parsing based on deep convolutional neural networks (DCNNs) have made significant progress. However, hierarchically exploiting multiscale and spatial contexts as convolutional features is still a hurdle to overcome. In order to boost the scale and spatial awareness of a DCNN, we propose two effective structures, named "Attention SPP and Attention RefineNet," to form a Mutual Attention operation, to exploit multiscale and spatial semantics different from the existing approaches. Moreover, we propose a novel Attention Guidance Network (AG-Net), a simple yet effective architecture without using bells and whistles (such as human pose and edge information), to address human parsing tasks. Comprehensive evaluations on two public datasets well demonstrate that the AG-Net outperforms the state-of-the-art networks.

1. Introduction

Human parsing, which segments a human image into the regions of semantic parts, has recently received considerable interest in computer vision areas. Due to its comprehensive and elaborate analysis of human information, human parsing has served as an indispensable basis for many high-level computer vision applications, such as objection detection [1], clothing parsing [2], human pose estimation [3–6], video surveillance [7, 8], and person reidentification [9].

A significant progress on human parsing has been achieved using a deep convolutional neural network (DCNN). However, the diversity of human poses, the foreshortening caused by viewpoint change, and the variance distribution of human bodies affect the accuracy of human parsing. For example, due to severe foreshortening and unusual pose, the upper-body of a human has a larger scale than the lower-body in Figure 1(a), and the shoes appear at the right side in Figure 1(b). The body part scales and body distributions of humans in Figure 1 are different from those in the majority of scenarios. Therefore, how to design a powerful and robust model to capture multiscale and spatial contextual information is crucial to address the human parsing task.

Confronted by the hurdle of exploiting multiscale features, some multibranch tactics have been proposed. The Spatial Pyramid Pooling (SPP) [10–12] and the RefineNet [13] approaches, where parallel convolution layers with different receptive fields are used to capture multiscale information, are two prevalent strategies to get over this hurdle. However, these multibranch methods simply employ only a concatenation or an additional operation to achieve a feature fusion, hence producing feature redundancies and suppressing the representation capacity of the whole network. Moreover, human parsing has an important characteristic different from other segmentation tasks. It greatly requires spatial awareness to parse spatial-oriented labels, such as right-arm, left-arm, right-shoe, and left-shoe. However, SPP and RefineNet are limited to capturing spatial semantics, because the multibranch methods have no special design to distinguish the upper and lower parts, and right and left parts of a human body, particularly in the challenging human images as shown in Figure 1.

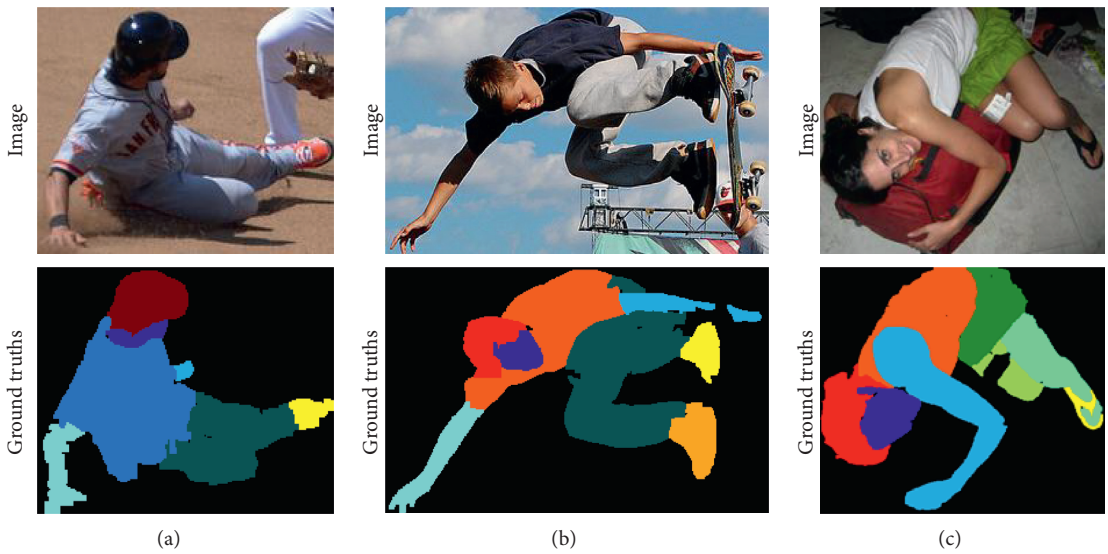

Figure 1: Illustration of challenging scales and spatial variations of real-world human images from the LIP dataset [3]. In (a), the scales of the upper-body and the head are larger than that of the lower-body. In (b), due to the transverse distribution of a body, the spatial locations of human parts are different from usual human images. In (c), both scale and spatial variations appear.

In order to address the challenges of multiscale and spatial feature extraction in human parsing, we impose a soft-attention mechanism into SPP and RefineNet methods to recalibrate high-level features in the model, hence producing the Attention SPP and Attention RefineNet, correspondingly. Based on the characteristic of multibranch model, we develop a light-weight trainable mechanism, named Mutual Attention (MA).

To exploit multiscale and spatial awareness with the attention-oriented philosophy, we propose an efficient Attention Guidance Network (AG-Net) for human parsing, which is shown as Figure 2. Specifically, the AG-Net can be divided into four steps. Firstly, we leverage fully convolutional network to encode a high-level feature map of the input image. Based on the high-level feature, secondly, we use the Attention SPP module capturing multiscale and spatial information. Thirdly, we decode the output feature from Attention SPP and each scale of decoding stage has a supervision strategy to supervise our model. Finally, each stage of decoding feature map is guided by Attention RefineNet to further fuse multiscale and spatial features.

Comprehensive experiments are conducted on two human parsing benchmark datasets, the ATR [14] and LIP [3] datasets, to evaluate our model. We demonstrate the feasibility and superiority of our methods on the ATR dataset.

Besides, we also use the ATR and LIP datasets to synthetically evaluate our AG-Net and obtain a state-of-the-art performance. In particular, in the evaluation of fine-grained and spatial-oriented labels, our approach obtains substantial improvement, which illustrates the remarkable ability of our AG-Net for human parsing.

There are two main contributions in our paper:

(i) To hurdle the issues of feature redundancies and spatial semantic limitations in SPP and RefineNet, we propose Attention SPP and Attention RefineNet and form Mutual Attention to recalibrate models

(ii) A portable and powerful architecture, named Attention Guidance Network (AG-Net), is designed to boost the multiscale and spatial semantic presentation ability in a deep learning model and obtain a brilliant human parsing performance

The remainder of this paper is organized as follows. In Section 2, we review related works. Subsequently, we describe each part of the proposed network in detail in Section 3. The experiments and conclusions are provided in Sections 4 and 5, respectively.

2. Related Works

Due to the great scientific value and commercial potential, human parsing has attracted increasing research interest [15–18] in recent years. In particular, significant progress on human parsing has been made using a fully convolutional network (FCN) [19]. However, the diversity and complication of real-world scenes make it hard to improve the accuracies of parsing results. Therefore, how to exploit the multiscale and spatial features is a key point to boost the parsing performance.

Recently, based on deep learning frameworks, there are two types of mainstream methods to improve the human parsing performance. The first type adopts some extra human body information to construct a model. The second type aims at exploiting the multiscale features of a human.

2.1. Introducing Extra Human Body Information. Introducing extra human body information (e.g., pose information or a structural relationship of different human body parts) aims at exploiting the spatial features of humans and improving the parsing results toward spatial-oriented labels. The methods of MuLA [6], LIPNet [3], and JMPP [4]

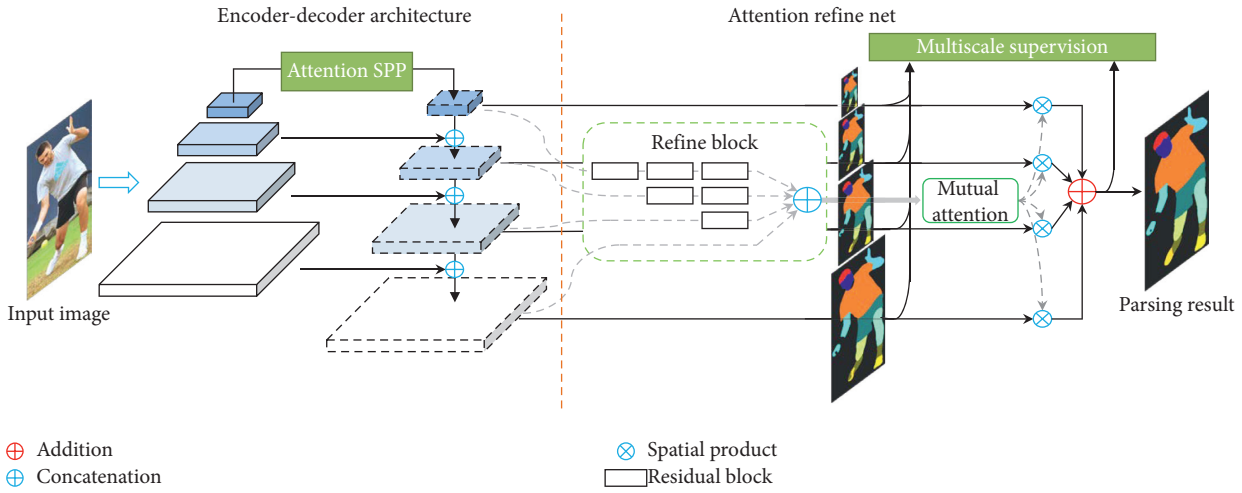

Figure 2: An overview of the Attention Guidance Network (AG-Net). Our network is established upon an Encoder-Decoder architecture extracting features from four resolution layers (i.e., 321×321, 161×161, 81×81, 41×41), and the Attention RefineNet learns the attention score maps to optimize the predicted label results. We impose the Attention SPP into the end of the encoder and leverage a multiscale supervision strategy to refine our model.

were established via combining human parsing and pose estimation into a network. With the pose information, the joint network could generate refined parsing results. PCNet [20] manually divided human body parts into different levels and established an end-to-end network to parse the human body parts from coarse to fine. Based on the ATR datasets and the CPC datasets, Guo et al. [21] applied prior pose information to increase the parsing accuracy. Different from the methods above, Luo et al. [15] proposed a trusted guidance learning framework to address the label parsing fragmentation issue. Su et al. [22] leveraged a label trusted network to solve the label confusion problem with the prior statistics of labels. Liu et al. [23] proposed a braiding network for fine-grained human parsing. At the end of this model, the semantic ambiguity of different body parts is eliminated with the help of pairwise hard regionpriors. Wang et al. [24] treated human parsing as a multisource information fusion process by combining convolutional neural network (CNN) with the compositional hierarchy of human bodies. Liu et al. [25] leveraged various priors such as feature resolution, global context information, and edge details to improve the human parsing performance. Huang et al. [26] rebalanced the imbalanced dataset from the perspective of geometry.

Although those methods could promise parsing performance, they required extra human body ground-truth information and hence greatly increased the workload of tagging datasets.

2.2. Exploiting the Multiscale Features. Exploiting the multiscale features is to obtain high-level abstract semantics without losing the semantic information of detailed texture synchronously. For example, DeepLab-v2 [10] created an ASPP model by employing the atrous convolution layer in a Spatial Pyramid Pooling structure. Inspired by the image pyramid, Chen et al. [17] trained several weight-sharing networks on different scales and merged the multiscale outputs into an attention network. In this way, high-level features with rich class information can be used to weight the underlying information to select details with precise resolutions. Co-CNN [18] imposed global image-level context and local super-pixel context into a unified model. To maintain pixel-level location information, Li et al. [27] and Chen et al. [28] used a pyramid structure to learn the attention mask instead of directly learning the feature map. During the decoding phase, this model introduced an attention mechanism that uses a high-resolution feature map to predict a channel mask. After that, this model multiplies the predicted channel mask with a low-resolution feature map shortcut. Huang et al. [29] proposed a novel trilateral awareness mechanism to sense the feature maps in trilateral levels to obtain comprehensive multiscale, spatial, and feature distribution information to exploit the semantic information precisely.

This type of methods greatly improves the accuracy of a deep model for human segmentation results, particularly with scale-orient labels. However, this method also greatly increases the number of parameters by dilating the structure of network. Moreover, it is extremely limited to extraction of position features to distinguish position-orient labels in human parsing tasks.

Different from the methods mentioned above, we propose Attention SPP and Attention RefineNet, which learn multiscale and spatial features simultaneously through Mutual Attention guidance. Moreover, an efficient AG-Net is proposed to address the challenges of human parsing. Compared with other methods, our model is simple yet effective in exploiting multiscale and spatial features without any bells and whistles (such as human pose or edge information).

3. Attention Guidance Network (AG-Net)

In this paper, based on the Encoder-Decoder architecture, we propose a novel Attention Guidance Network (AG-Net)

for human parsing task, as shown in Figure 2. In the end of the encoder part and each scale of decoder part, we impose the Attention SPP and Attention RefineNet correspondingly. Guided by Mutual Attention, the SPP and RefineNet have further powerful capacity to exploit multiscale and spatial semantics. Therefore, the whole model is designed with the attention-guided philosophy, which aims at selectively emphasising informative features and restraining less useful ones, and then the network has much powerful awareness to handle the complicated multiscale- and spatial-oriented features in human parsing task.

3.1. Mutual Attention (MA). The structure of Mutual Attention (MA), as shown in Figure 3, is composed of a Spatial Attention and a Channel Attention part. The Spatial Attention part concentrates on optimizing position sensitive features such as the location distribution of a human pose and organ, and it enhances the spatial perception and generalization ability of the model. Using the Channel Attention optimizes the cross-channel contexts due to emboldening informative semantics and dampening valueless one. Therefore, MA achieves the goal of recalibration of the position and channel contexts in a feature matrix.

Let X and Y be the input and output feature matrices of $C \times H \times W$ dimensions, where H and W denote the spatial dimensions and C denotes the channel. For feature extraction in the Conv block $F(\cdot)$, the thickness of output feature matrix $f_i^c = F(X)$ is C, where i represents the feature points in spatial dimensions. Through max pooling of stride = 1 by a 3×3 filter, two 3×3 convolutions, and a softmax operation, a Spatial Attention map w_i is produced. A ReLU [31] is embedded in the two convolutions. The matrix s_i^c generated by the Spatial Attention operation is defined as

$$s_i^c = w_i \cdot f_i^c. \tag{1}$$

Using the average global pooling approach, the s_i^c is transferred to a feature vector V^C, whose element v^c is calculated by

$$v^c = \frac{1}{W \times H} \sum_{i=1}^{W \times H} s_i^c. \tag{2}$$

In order to reduce the feature parameters, referred to as the SENet [32], we employ a squeeze and excitation structure to extract channel level features. Due to the sharp compression of channel features in hourglass structure, if a nonlinear activation function, such as the ReLU function, is introduced, some useful informative features will be inevitably lost [33]. Therefore, distinguished from the SENet, we do not employ the ReLU operation after its squeezing FC layer so as to preserve the integrity of useful contexts. The output of Channel Attention vector is as follows:

$$z^c = \sigma(W, v^c) = \sigma(W_2(W_1 v^c)), \tag{3}$$

where σ denotes the sigmoid function. W_1 and W_2 are the weights of two FC layers. By the Channel Attention part, the final output of the features Y is defined as

$$Y = z^c \cdot s^c. \tag{4}$$

Mutual Attention can rebalance the features in the spatial level and channel level with negligible parameters increasing. From the perspective of attention, this mechanism achieves the goal of feature decoupling and boosts the capacity of feature expression in the network.

3.2. Attention SPP. The SPP uses the multiscale convolution filter and pyramid pooling structure to extract a feature pyramid from high-level semantics. Due to its powerful performance, SPP has been imposed into many semantic segmentation tasks. However, because of the massive size of its multibranch fusion model, SPP unavoidably has redundant and overcoupling feature information in its output, so that the features cannot be fully utilized and even the network has to be retrained or relearned. Moreover, SPP is lacking in spatial semantics capturing, which cannot satisfy the need of parsing the spatial-oriented labels.

In order to improve feature utilization for multiscale feature mining and impose the spatial feature extraction ability in SPP, we propose the Attention SPP model. We illustrate the specific design of Attention SPP shown in Figure 4, which uses the Atrous Spatial Pyramid Pooling (ASPP) [28] as an example. Confronted by the different branches in SPP, at the spatial level, Spatial Attention is employed to rebalance different-scale semantics correspondingly to enable the branches to concentrate on their own jobs. Besides, using the Spatial Attention can guide the convolution layers to focus on reweighting the position and angular distributions based on multiscale receptive fields; it enables the SPP to parse the spatial semantics. In the end of model, we use the concatenation operation to aggregate all paths and generate a feature matrix with affluent contextual information. Nevertheless, with the sparse and redundant features, the output matrix will hinder the feature extraction of latter layers, thus dampening the presentational capability of the network. Consequently, the Channel Attention is employed to recalibrate the feature-rich but redundant matrix in the channel level.

Finally, through using the self-learning weight coefficients to rebalance the matrix, the output features contain much more abundant multiscale contextual information. Additionally, the SPP model is efficient in capturing spatial features. When the multiscale and spatial feature extraction intertwine in one model, the Attention SPP has much more representational power to confront human parsing tasks.

3.3. Attention RefineNet. To absorb the multiscale and exploit spatial contextual information generated by Encoder-Decoder architecture with spatial information, we impose the Attention RefineNet into the AG-Net, shown in Figure 2, which can then further optimize the predicted score maps.

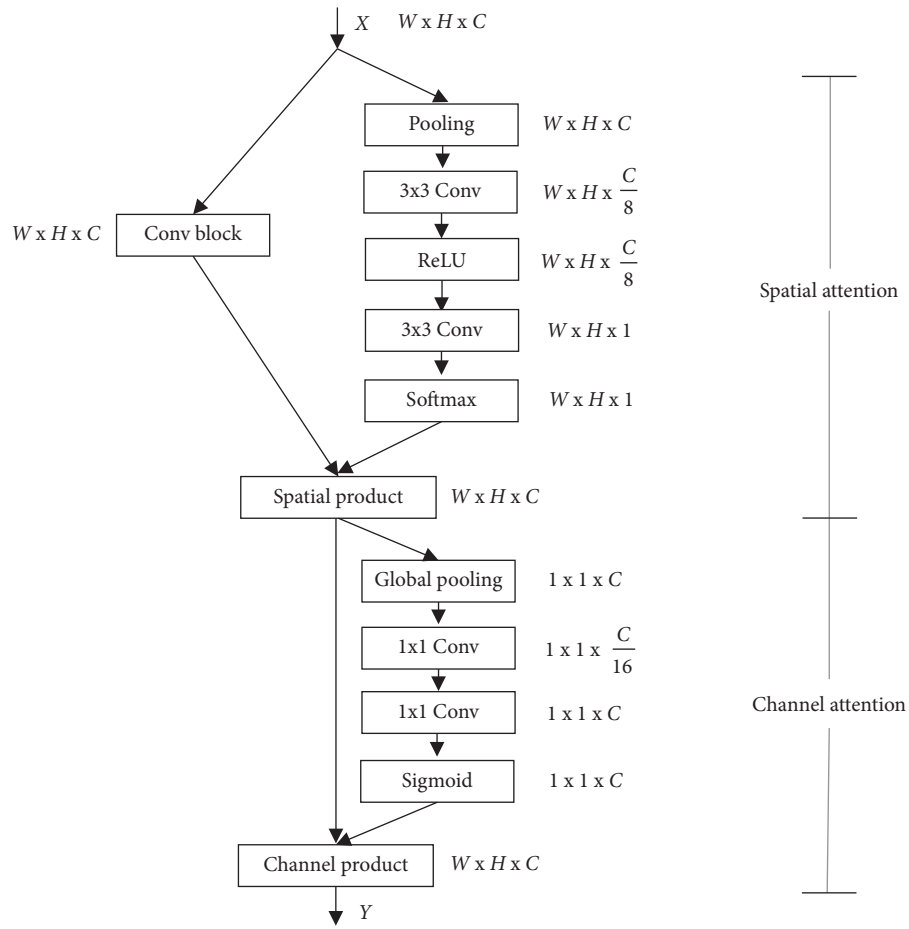

Figure 3: The design of Mutual Attention (sequential type). The Conv block is not a specifically defined convolution structure, and it can be defined as a Residual [30] block, ASPP [10], or others. In our strategy, the input and output feature maps must have a uniform resolution.

Inspired by CPNet [34], confronted by downsampled (2, 4, 8) feature layers, we perform a series of operations to cascade 1, 2, and 3 Residual blocks, respectively. In the end of the Refine block, we integrate the information of different levels via a concatenating operation. In order to reduce the complexity of the model, a bilinear interpolation is used to unify the multiscale features to of 161×161 resolution instead of the original input scale.

Like SPP, the hurdle caused by the issues of feature redundancies and spatial context limitations also adheres to the feature matrix generated by multiscale feature fusions. Accordingly, MA is injected into RefineNet, which rebalances the feature matrix at the spatial level and channel level. MA makes different paths of RefineNet focus on their corresponding-scale features and explore the potential of spatial feature extraction. In addition, it also redistributes the learning direction of the model, reduces the learning difficulty of the original task, and makes the network easier to train.

3.4. Supervision Strategy. There have been existing works [34, 35] using a multiscale supervision strategy to refine their models. The proposed AG-Net also adopts this strategy as shown in Figure 2.

Based on the traditional softmax loss function in pixel-wise ground-truth masks, we inject a global loss function as follows to optimize AG-Net:

$$L = L_{\text{mask}}(P, G) + \lambda L_{\text{global}}(V_P, V_{G.T.}), \tag{5}$$

where P is a predicted label map and G is the GT mask. V_P denote the global predict vector extracted from the predicted label map using global pooling operation, and $V_{G.T.}$ denotes the global GT vector extracted from GT mask. L_{mask} represents the common loss function of semantic parsing, and λ is the weight of global loss L_{global}, which is obtained by

$$L_{\text{global}} = \frac{1}{2N} \sum_{n=0}^{N-1} \left\| v_{\text{gt}}^n - v_P^n \right\|_2^2, \tag{6}$$

where N denotes the count of ground-truth labels. In the experiments of our model, the hyperparameter λ is set to 0.1.

4. Experiment Analysis

In the training process, we remove the softmax and full connected layers of the VGG-16 [36] and DenseNet-121 [37] and replace them with a fully convolutional network to extract features. Besides, we utilize the hybrid dilated convolution in the conv5 layers. The input image is of 321×321

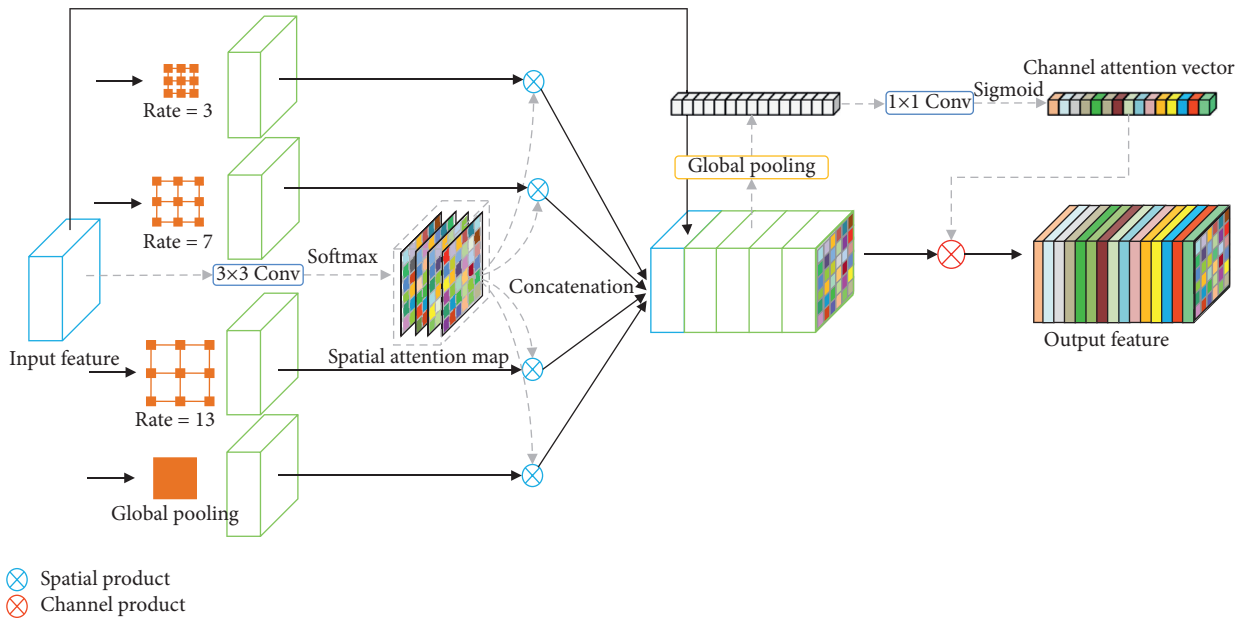

Figure 4: The structure of Attention SPP. Based on input features of 41 × 41 resolution in AG-Net, we design three atrous branches with rates of 3, 7, and 13, respectively, and a global pooling branch to constitute our model. Besides, exploited from the input features, we propose a Spatial Attention branch to highlight the corresponding multiscale and spatial semantics. After the concatenation operation, we leverage a Channel Attention branch to optimize the feature-rich but redundant matrix in the channel level. The gray links indicate the flow of Mutual Attention operation imposed in SPP.

cropped from an original image with about 10 FPS in the VGG-16 based model and 15 FPS in the DenseNet-121 based model. We adopt the initialization with a pretrained model and leverage the Gaussian distribution with standard deviation of 0.01 to initialize each without pretrained layers. We utilize the Adam [38] solver with batch size of 6, momentum of 0.9, weight decay of 0.0005, and initial learning rate of 0.0001. Inspired by the DeepLab method [10], we use the poly strategy $(1 - (\text{iter}/\text{max\_iter}))^{\text{power}}$ to dynamically adjust the learning rate. The training data are augmented by a left-right flipping. All models are experienced using the PyTorch platform [39]. Our experiments are implemented on a system of Core Intel i7-5930K CPU and a single NVIDIA GTX 1080 Ti GPU. We estimate experimental results on two human parsing benchmark datasets, the LIP [3] and ATR [14] datasets. We train our model with 30 epochs and 60 epochs in ATR and LIP, respectively.

4.1. Datasets

4.1.1. LIP Dataset. The Look into Person (LIP) dataset contains 50,462 images with careful pixel-wise annotations of 19 semantic human parts from the MS COCO dataset. These images are collected from real-world scenes and present various and complicated views of human appearances, poses, sizes, clothes, occlusion, illumination, and feature confusion. Besides, it not only categorizes the traditional human parts, but also annotates some tiny labels (e.g., sunglasses, socks). Some annotations contain spatial-oriented information (e.g., left-arm, right-arm, left-shoe, right-shoe). Therefore, it is a challenging human parsing dataset.

4.1.2. ATR Dataset. ATR dataset contains a total of 17,700 images, which consist of 7,700 images in the original ATR [14] and 10,000 additional images in the Chictopia10K [18]. All images are annotated pixel-wise with 18 categories. For a convenient comparison, we follow the setting of Co-CNN [18] and split the original ATR dataset into 700 images for validation, 1,000 images for testing, and the rest for training.

We only experiment on these two datasets because other datasets do not have fine-grained and spatial-oriented segment annotations to satisfy the needs of our model.

4.2. Attention SPP Evaluation and Discussion. To verify the advantages of our Attention SPP, we embed our MA into four classic SPP methods [10–12]: DeepLab v2, DeepLab v3, Vortex, and PSPNet. We evaluate these four methods on the ATR benchmark with the same experiment settings. All models are trained with the VGG-16 bottleneck for 30 epochs. We deploy two effective measurements, the average F1 score and the mean intersection-over-union (mIoU), to compare the performance of the models. We can see from Table 1 that each model can be increased by 1 ~ 3% on both F1 score and mIoU with MA. The Vortex method with the sequential type of MA, especially, is increased by 2.94% and 4.42% in terms of F1 score and mIoU, respectively.

We depict testing curves of four models with/without Mutual Attention (MA) on the ATR dataset in Figure 5. For each model, both types of MA can improve model performance and accelerate the convergence. It is demonstrated that the attention structure can enhance the representative and comprehensive ability of the models by guiding the networks to capture more useful information.

Table 1: Embedding two types of Mutual Attention (MA) into four classical SPP models to evaluate them on the ATR test set.

Method	Avg. F1 score	mIoU
DeepLab-v2 [10]	73.21	58.50
DeepLab-v2 + MA (parallel)	73.36	61.01
DeepLab-v2 + MA (sequential)	73.94	61.43
DeepLab-v3 [28]	73.34	59.41
DeepLab-v3 + MA (parallel)	75.43	62.41
DeepLab-v3 + MA (sequential)	75.44	62.39
PSPNet [11]	73.46	59.27
PSPNet + MA (parallel)	75.27	62.15
PSPNet + MA (sequential)	75.41	62.70
Vortex [12]	72.96	58.55
Vortex + MA (parallel)	75.32	62.16
Vortex + MA (sequential)	75.49	62.97

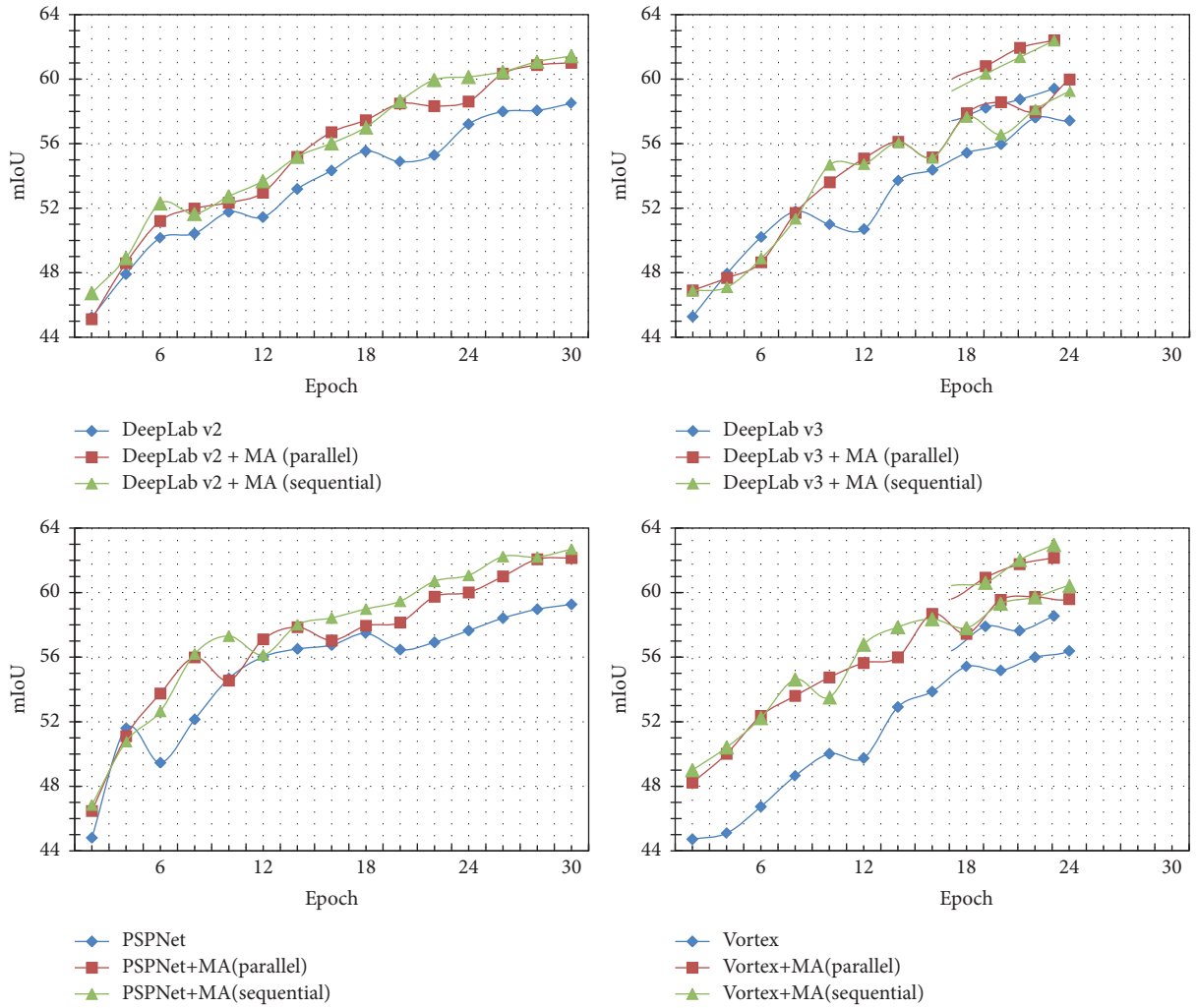

Figure 5: Testing curves of DeepLab-v2, DeepLab-v3, PSPNet, and Vortex methods with/without Mutual Attention (MA) on ATR.

Furthermore, the MA employing the sequential operation outperforms the parallel operation on aggregating the spatial and Channel Attention parts, as shown in Table 1. The sequential operation achieves the feature decoupling purpose. The Spatial Attention focuses on learning spatial contexts. Then, based on the optimized spatial contexts, the Channel Attention attentively recalibrates cross-channel contexts. This sequential structure can dynamically guide the model to use features of different levels in different stages, so as to alleviate the difficulty of a deep learning.

Due to the best performance of Vortex method [12] with the sequential type of MA in Table 1, we deploy the Vortex method, which leverages the Vortex Pooling strategy to improve the feature utilization ratio in the original SPP, for our Attention SPP backbone into AG-Net.

4.3. Human Parsing Performances. We evaluate the comprehensive performance of the proposed AG-Net for human parsing in ATR and LIP datasets.

4.3.1. ATR Dataset. Table 2 shows the result of the proposed method with ten state-of the-art methods on the ATR dataset. Due the evaluation settings in Co-CNN [18] and TGPNet [15], we employ the average F1 score as the evaluation criteria. From Table 2, we can see that both types of our AG-Net have achieved remarkable results. Compared with the state-of-the-art approaches, our method with the backbone of DenseNet-121 has seen an improvement of 0.5%.

4.3.2. LIP Dataset. According to the evaluation method introduced in LIP [3], Table 3 shows comparison results with fifteen state-of-the-art approaches with mIoU measurements and Table 4 shows eight state-of-the-art approaches with IoU measurements on 20 class labels.

In Table 3, the Baseline network means the combination of the Encoder-Decoder architecture, Vortex bottleneck, and multiscale supervision. On the strength of the Baseline, we, respectively, inject the Attention SPP or Attention RefineNet into the model, making 1.90% and 3.37% improvements compared with the Baseline approach. Additionally, AG-Net with VGG-16 backbone enormously improves by 5.92% compared with the Baseline. However, the network scale only expands a little (Baseline 156M vs. AG-Net 161M) and the training speed remains almost identical. Toward the backbone of DenseNet-121, our AG-Net outperforms all state-of-the-art methods with relatively small network size (106M). We use only the traditional pixel-wise mask supervision instead of adding auxiliary pose labels to supervise our model. Therefore, our model is a simple and efficient model to conduct human parsing tasks.

Five methods (i.e., CE2P [25], BraidNet [23], HRNetV2 [46], CNIF [24], A-CE2P [47]) obtain very high scores by introducing various priors such as human body can be represented as a hierarchy of multilevel parts or others. Nevertheless, these priors require introducing additional datasets [46] and additional networks [24, 25, 46], which lead to complex and inefficient network structures, as well as extra costs of tagging and training. In addition, some of these models have evolved into multistage inference approaches [23, 24, 46], which are not flexible enough to be embedded in other tasks.

Our DenseNet-121 based model achieves the best performance on 13 mIoU results with detailed labels as shown in Table 4. Note that for the labels that require high-level spatial-oriented features to distinguish the global direction and spatial position, such as left-arm, right-arm, left-leg, right-leg, left-shoe, and right-shoe, our model can surpass the existing state-of-the-art approaches by 6.83%, 5.05%, 4.74%, 6.43%, 9.49%, and 6.56%, respectively. Besides, our proposed AG-Net also gets the best results on scale-oriented labels, such as gloves, socks, and hat, which all need fine-grained features to guide. Moreover, based on the same network settings with VGG-16 backbone, the AG-Net can significantly improve the Baseline on gloves, sunglasses, socks, skirt, right- and left-leg, and right- and left-shoe by about 10%. The great improvement demonstrates the strong generalization ability of our network in extracting multiscale and spatial features for human parsing.

Figure 6 shows the Spatial Attention heatmaps toward different scales of receptive fields in the Attention SPP and Attention RefineNet. For the layers with small receptive fields, the Spatial Attention heatmaps tend to focus on the edges and small parts while large receptive fields cause the

Table 2: Human parsing performances with our network and ten state-of-the-art methods on ATR [14] test set.

Method	Avg. F1 score
Yamaguchi et al. [40]	41.80
Paper Doll [2]	44.76
M-CNN [41]	62.81
ATR [14]	64.38
DeepLab-v2 (VGG16) [10]	73.53
Attention (VGG16) [17]	77.23
DeepLab-v3 + [42]	76.97
Co-CNN-v1 [18]	80.14
TGPNet [15]	81.76
Co-CNN-v2 [43]	85.36
AG-Net (VGG16)	83.59
AG-Net (DenseNet-121)	85.46

Table 3: Human parsing results with fifteen state-of-the-art methods on the LIP validation set.

Method	mIoU
SegNet [44]	18.17
FCN-8s [19]	28.29
DeepLab (VGG16) [10]	41.64
Attention [17]	42.92
DRN-50 + Vortex [45]	41.09
DeepLab (ResNet-101) [10]	44.80
SS-JPPNet [3]	44.73
SS-NAN [16]	47.92
SPReID [9]	48.16
MuLA [6]	49.30
CE2P [25]	53.10
BraidNet [23]	54.40
HRNetV2 [46]	56.48
CNIF [24]	57.74
A-CE2P [47]	59.36
Baseline (VGG16)	40.41
Baseline + Attention ASPP	42.31
Baseline + Attention RefineNet	43.78
AG-Net (VGG16)	46.33
AG-Net (DenseNet-121)	50.54

The Baseline is AG-Net without the attention mechanism proposed in this paper.

Table 4: Per-label comparison of per-class IoU with eight state-of-the-art methods on the LIP validation set.

Method	SegNet	FCN-8s	DeepLab (VGG16)	Attention	DRN50 + Vortex	DeepLab (Res-101)	SS-JPPNet	SS-NAN	Baseline (VGG16)	AG-Net (VGG16)	AG-Net (Dense-121)
Hat	26.60	39.79	57.94	58.87	57.50	59.76	59.75	63.86	58.06	63.27	66.24
Hair	44.01	58.96	66.11	66.78	67.37	66.22	67.25	70.12	68.14	69.52	71.25
Gloves	0.01	5.32	28.50	23.32	13.55	28.76	28.95	30.63	12.50	23.92	38.74
Sunglasses	0.00	3.08	18.40	19.48	18.82	23.91	21.57	23.92	18.96	29.59	30.45
U-clothes	34.46	49.08	60.94	63.20	62.52	64.95	65.30	70.27	62.33	64.62	67.12
Dress	0.00	12.36	23.17	29.63	23.88	33.68	29.49	33.51	22.49	26.37	31.06
Coat	15.97	26.82	47.03	49.70	48.51	52.86	51.92	56.75	45.12	47.31	53.56
Socks	3.59	15.66	34.51	35.23	33.64	37.67	38.52	40.18	30.72	41.02	45.69
Pants	33.56	49.41	64.00	66.04	65.79	68.05	68.02	72.19	64.98	68.98	73.34
Jumpsuit	0.01	6.48	22.38	24.73	14.10	26.15	24.48	27.68	15.66	20.52	25.52
Scarf	0.00	0.00	14.29	12.84	0.84	17.44	14.92	16.98	8.05	11.75	17.22
Skirt	0.00	2.16	18.74	20.41	19.41	25.23	24.32	26.41	12.31	22.65	27.60
Face	52.38	62.65	69.70	70.58	71.99	70.00	71.01	75.33	71.92	73.40	74.34
L-arm	15.30	29.78	49.44	50.17	52.17	50.42	52.64	55.24	51.57	58.03	62.06
R-arm	24.23	36.63	51.66	54.03	55.25	53.89	55.79	58.93	54.82	60.35	63.98
L-leg	13.82	28.12	37.49	38.35	40.23	39.36	40.23	44.01	38.03	46.35	48.75
R-leg	13.17	26.05	34.60	37.70	37.63	38.27	38.80	41.87	37.85	45.18	48.30
L-shoe	9.26	17.76	28.22	26.20	27.58	26.95	28.08	29.15	25.67	32.88	38.64
R-shoe	6.47	17.70	22.41	27.09	26.53	28.36	29.03	32.64	25.13	35.03	39.20

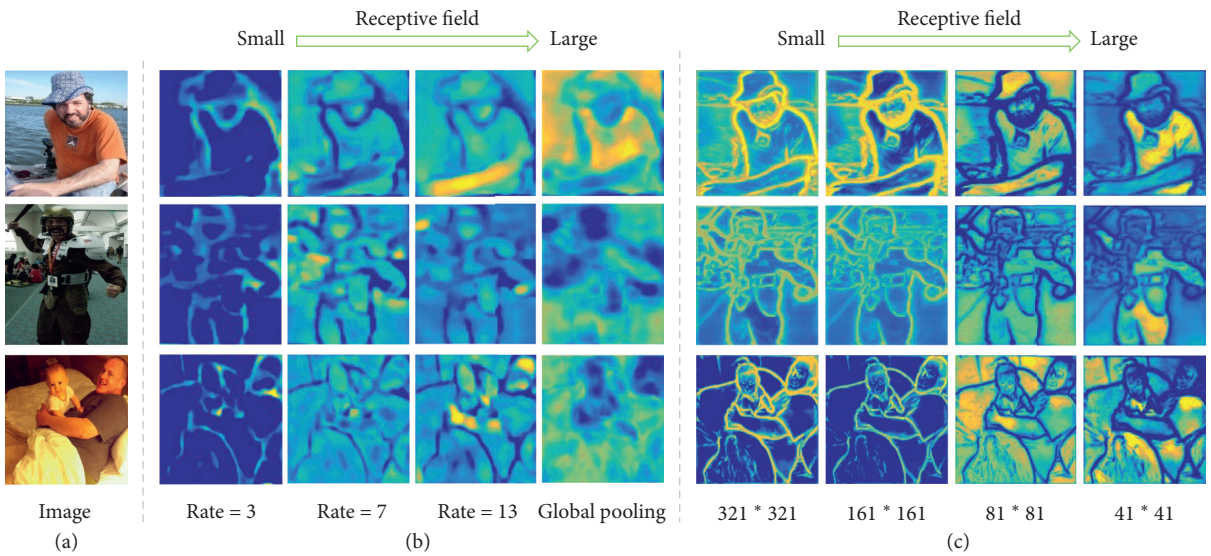

Figure 6: Spatial Attention heatmaps from the Attention SPP and Attention RefineNet toward different-scale receptive fields. The Attention RefineNet is much closer to the output of model, and the Attention SPP contains much high-level abstract information. Therefore, the heatmaps of the Attention RefineNet have more detailed texture than those of the Attention SPP. (a) Input image. (b) Spatial heatmaps from Attention SPP. (c) Spatial heatmaps from Attention RefineNet.

heatmaps to comprehend global information. For example, as shown in the first row of images in Figure 6(b), the heatmap with atrous rate = 3 has higher confidence on human edges while the heatmap with atrous rate = 13 tends to represent midlevel objects, such as hands, which need to distinguish the orientations. Global pooling, additionally, can better be aware of large scale regions, such as upper-clothes and foreground-background. To sum up, the Attention SPP and Attention RefineNet can achieve effective decoupling and can guide networks to represent multiscale features by reducing redundancies in different features. Besides, it can endow the model sensitivity toward the human parts and positions.

Visualized comparisons on the LIP dataset are shown in Figure 7. We compare the visual results of our AG-Net with three state-of-the-art methods, which are DeepLab (VGG16) [10], DRN−50 + Vortex [12, 45], and SS-JPPNet [3], and a Baseline network, on ten representative and challenging

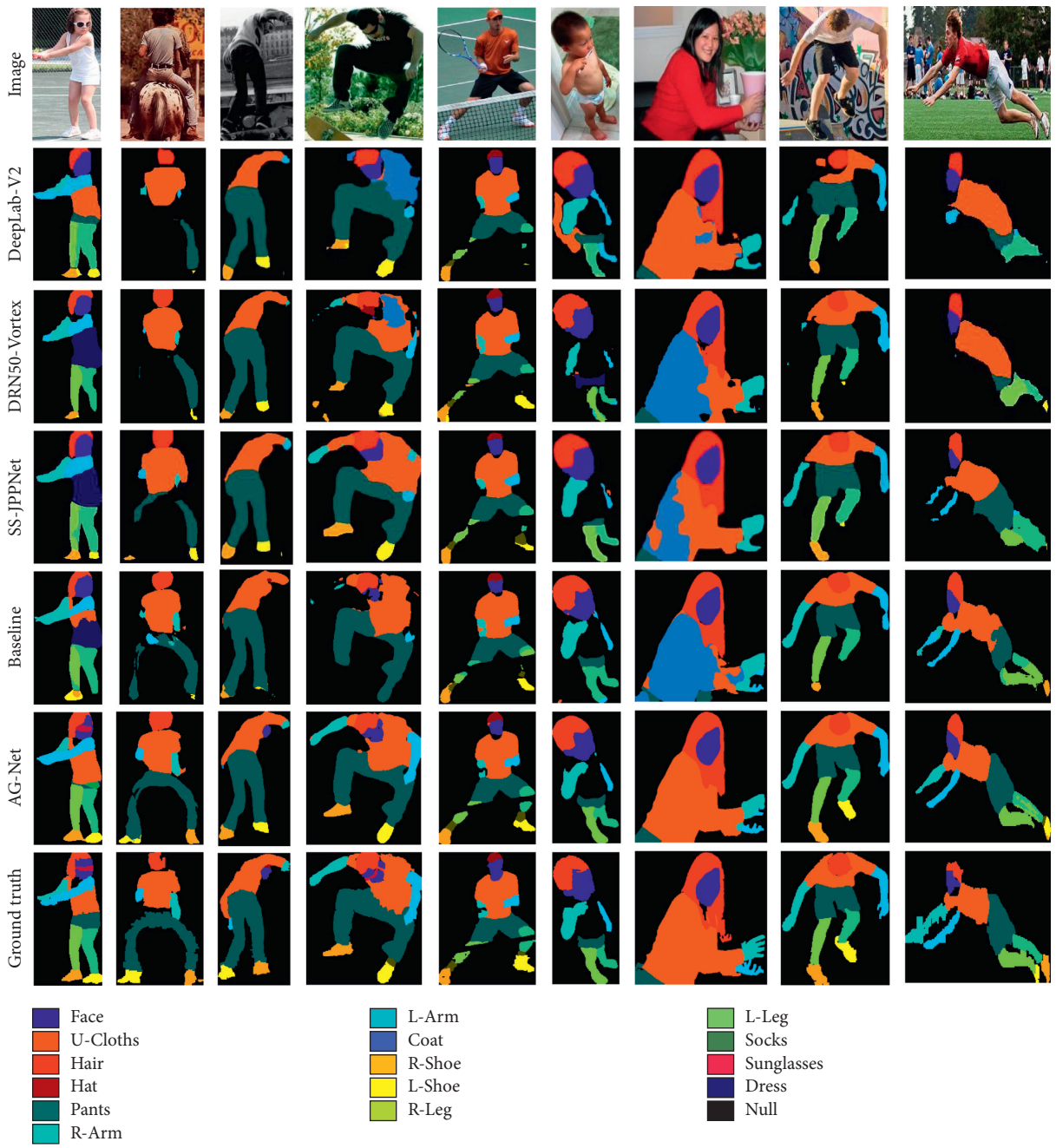

Figure 7: Visual comparison of human parsing results on the LIP validation set. In the images in columns 3 and 9, the ground truths wrongly annotate the right-shoes and left-shoes.

images. From columns 1 and 3, we can draw a conclusion that our AG-Net has a strong ability to segment small and confusing parts. The proposed AG-Net is the only method that successfully parses the sunglasses and head labels, respectively, in columns 1 and 3. For spatial-oriented labels, our AG-Net can accurately classify them without being influenced by the shooting angles and pose variations. For example, in column 6, the child has larger scale of upper-body than that of lower-body due to the high angles and side shots of images. Besides, in column 9, the human pose is different from those in commonly seen images and has a transverse distribution, while our AG-Net can still accurately get the paring results. Therefore, our method shows a robust performance in segmenting multiscale- and spatial-oriented human parts.

5. Conclusion

In this paper, we have proposed novel Attention SPP and Attention RefineNet and used a Mutual Attention

mechanism to recalibrate feature maps in bilateral levels (the spatial dimension and channel dimension) to get comprehensive multiscale and spatial features different from the existing approaches. Moreover, the Attention Guidance Network (AG-Net), a simple and efficient model designed with the attention-centric theory, has been proposed to boost human parsing performance in scale- and spatial-oriented labels. Extensive experiments on two benchmarks for human parsing have demonstrated the representational power of AG-Net and shown that our method outperforms the state-of-the-art methods.

Data Availability

The data used to support the findings of this study are included within the article.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This research was supported by the National Natural Science Foundation of China (61872394, 61672547, and 61772140), Guangzhou Science and Technology Plan Project (201902010056), and Guangxi Innovation Driven Development Special Fund Project (AA18118039).

References

[1] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask R-CNN," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pp. 2980–2988, Venice, Italy, October 2017.

[2] K. Yamaguchi, M. H. Kiapour, L. E. Ortiz, and T. L. Berg, "Retrieving similar styles to parse clothing," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 37, no. 5, pp. 1028–1040, 2015.

[3] X. Liang, K. Gong, X. Shen, and L. Lin, "Look into person: joint body parsing & pose estimation network and a new benchmark," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, no. 4, pp. 871–885, 2018.

[4] F. Xia, P. Wang, X. Chen, and A. Yuille, "Joint multi-person pose estimation and semantic part segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6080–6089, Honolulu, HI, USA, July 2017.

[5] X. Nie, J. Feng, Y. Zuo, and S. Yan, "Human pose estimation with parsing induced learner," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2100–2108, Salt Lake City, Utah, June 2018.

[6] X. Nie, J. Feng, and S. Yan, "Mutual learning to adapt for joint human parsing and pose estimation," in *Proceedings of the European Conference on Computer Vision (ECCV)*, Munich, Germany, September 2018.

[7] S. Liu, Y. Sun, D. Zhu et al., "Cross-domain human parsing via adversarial feature and label adaptation," in *Proceedings of the AAAI Conference On Artificial Intelligence*, New Orleans, LA, USA, February 2018.

[8] Y. Han, G. RenR. Qian et al., "A real-time surveillance video parsing system with single frame supervision," in *Proceedings of the 2017 ACM on Multimedia Conference*, pp. 1257-1258, ACM, Mountain View, CA USA, October 2017.

[9] M. M. Kalayeh, E. Basaran, M. Gokmen, M. E. Kamasak, and M. Shah, "Human semantic parsing for person re-identification," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Salt Lake City, UT, USA, June 2018.

[10] 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 Transactions on Pattern Analysis and Machine Intelligence*, vol. 40, no. 4, pp. 834–848, 2018.

[11] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6230–6239, Honolulu, HI, USA, July 2017.

[12] C. W. Xie, H. Y. Zhou, W. Lu, and J. Wu, "Vortex pooling: improving context representation in semantic segmentation," in *Proceedings of the 2019 IEEE 31st International Conference on Tools with Artificial Intelligence (ICTAI)*, Portland, OR, USA, November 2018.

[13] G. Lin, A. Milan, C. Shen, and I. Reid, "Refinenet: multi-path refinement networks for high-resolution semantic segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Honolulu, HI, USA, July 2017.

[14] X. Liang, S. Liu, X. Shen et al., "Deep human parsing with active template regression," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 37, no. 12, pp. 2402–2414, 2015.

[15] X. Luo, Z. Su, J. Guo, G. Zhang, and X. He, "Trusted guidance pyramid network for human parsing," in *Proceedings of the 2018 ACM on Multimedia Conference*, Seoul, Korea, October 2018.

[16] J. Zhao, J. Li, X. Nie et al., "Self-supervised neural aggregation networks for human parsing," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1595–1603, Honolulu, HI, USA, July 2017.

[17] 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 (CVPR)*, pp. 3640–3649, IEEE, Las Vegas, NV, USA, June 2016.

[18] X. Liang, C. Xu, X. Shen et al., "Human parsing with contextualized convolutional neural network," in *Proceedings of the International Conference on Computer Vision (ICCV)*, pp. 1386–1394, Santiago, Chile, December 2015.

[19] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," *IEEE Transactions On Pattern Analysis And Machine Intelligence*, vol. 39, pp. 640–651, 2017.

[20] B. Zhu, Y. Chen, M. Tang, and J. Wang, "Progressive cognitive human parsing," *AAAI Conference On Artificial Intelligence*, vol. 34, no. 7, 2018.

[21] J. Guo, Z. Su, X. Luo, G. Zhang, and X. Liang, "Conditional feature coupling network for multi-persons clothing parsing," in *Proceedings of the Advances in Multimedia Information Processing–PCM 2018,Lecture Notes in Computer Science*, pp. 189–200, Springer, Hefei, China, September 2018.

[22] Z. Su, J. Guo, G. Zhang, X. Luo, R. Wang, and F. Zhou, "A conditional progressive network for clothing parsing," *IET Image Processing*, vol. 13, no. 4, pp. 556–565, 2018.

[23] X. Liu, M. Zhang, W. Liu, J. Song, and T. Mei, "Braidnet: braiding semantics and details for accurate human parsing,"

in *Proceedings of the 27th ACM International Conference on Multimedia*, pp. 338–346, Nice, France, October 2019.

[24] W. Wang, Z. Zhang, S. Qi, J. Shen, Y. Pang, and L. Shao, "Learning compositional neural information fusion for human parsing," in *Proceedings of the International Conference on Computer Vision (ICCV)*, pp. 5703–5713, Seoul, Korea, October 2019.

[25] T. Liu, T. Ruan, Z. Huang et al., "Devil in the details: towards accurate single and multiple human parsing," in *Proceedings of the AAAI Conference On Artificial Intelligence*, Honolulu, HI, USA, January 2019.

[26] E. Huang, Z. Su, F. Zhou, and R. Wang, "Learning rebalanced human parsing model from imbalanced datasets," *Image and Vision Computing*, vol. 99, Article ID 103928, 2020.

[27] H. Li, P. Xiong, J. An, and L. Wang, "Pyramid attention network for semantic segmentation," in *Proceedings of the British Machine Vision Conference (BMVC)*, London, UK, January 2018.

[28] L. C. Chen, G. Papandreou, F. Schroff, and H. Adam, *Rethinking Atrous Convolution for Semantic Image Segmentation*, Cornell University, Ithaca, NY, USA, 2017.

[29] E. Huang, Z. Su, and F. Zhou, "Tao: a trilateral awareness operation for human parsing," in *Proceedings of the IEEE International Conference on Multimedia and Expo (ICME)*, pp. 1–6, London, UK, June 2020.

[30] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, Las Vegas, NV, USA, July 2016.

[31] G. Nair and G. Hinton, "Rectified linear units improve restricted Boltzmann machines," in *Proceedings of the 24th international conference on Machine learning -ICML '07*, pp. 807–814, New York, NY, USA, June 2010.

[32] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Salt Lake City, UT, USA, June 2018.

[33] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L. Chen, "MobileNetV2: inverted residuals and linear bottlenecks," in *Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4510–4520, Salt Lake City, UT, USA, 2018.

[34] Y. Chen, Z. Wang, Y. Peng, Z. Zhang, G. Yu, and J. Sun, "Cascaded pyramid network for multi-person pose estimation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Salt Lake City, UT, USA, June 2018.

[35] M. A. Islam, M. Rochan, N. D. B. Bruce, and Y. Wang, "Gated feedback refinement network for dense image labeling," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4487–4885, Honolulu, HI, USA, October 2017.

[36] A. Simonyan and Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Las Vegas, NV, USA, June 2014.

[37] G. Huang, Z. Liu, L. V. D. Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2261–2269, Honolulu, HI, USA, July 2017.

[38] D. Kingma and J. Ba, "Adam: a method for stochastic optimization," in *Proceedings of the 3rd International Conference on Learning Representations (ICLR 2015)*, San Diego, CA, USA, May 2015.

[39] P. Adam, G. Sam, C. Soumith et al., "Automatic Differentiation in Pytorch," in *Proceedings of the Conference on Neural Information Processing Systems*, Long Beach, CA, USA, December 2017.

[40] K. Yamaguchi, M. H. Kiapour, L. E. Ortiz, and T. L. Berg, "Parsing clothing in fashion photographs," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3570–3577, Rhode Island, CA, USA, June 2012.

[41] X. Liu, L. Liang, X. Liu et al., "Matching-CNN meets KNN: quasi-parametric human parsing," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1419–1427, Boston, MA, USA, April 2015.

[42] L. C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," in *Proceedings of the Computer Vision–ECCV 2018,Lecture Notes in Computer Science*, Munich, Germany, October 2018.

[43] X. Liang, C. Xu, X. Shen et al., "Human parsing with contextualized convolutional neural network," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 1, pp. 115–127, 2017.

[44] V. Badrinarayanan, A. Kendall, and R. Cipolla, "SegNet: a deep convolutional encoder-decoder architecture for image segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 12, pp. 2481–2495, 2017.

[45] F. Yu, V. Koltun, and T. Funkhouser, "Dilated residual networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 636–644, Honolulu, HI, USA, May 2017.

[46] J. Wang, K. Sun, T. Cheng et al., "Deep high-resolution representation learning for visual recognition," in *Proceedings of the IEEE Transactions on Pattern Analysis and Machine Intelligence*, p. 1, Washington, DC, USA, April 2020.

[47] P. Li, Y. Xu, Y. Wei, and Y. Yang, "Self-correction for human parsing," in *Proceedings of the IEEE International Conference on Computer Vision Workshop*, Seoul, Korea, October 2019.