Human Part Semantic Segmentation Using Custom-CDGNet Network

Aditi Verma, Vivek Tiwari, Mayank Lovanshi, Rahul Shrivastava, and Basant Tiwari

Abstract Human body part segmentation is a semantic segmentation of human images task that entails labelling pixels in an image into their respective classes. The human body is composed of hierarchical structures in which each body part in the image has a particular individual location. Considering this knowledge, the sample class distribution technique was developed by collecting and applying the primary human parsing labels in vertical and horizontal dimensions. The proposed network exploits the underlying position distribution of the classes to make precise predictions with the help of these classes. We produce a distinct spatial guidance map by combining these guided features. This guidance map is then superimposed on our backbone network. Extensive experiments were executed on a large data set, i.e. LIP, and evaluation was done using the mean IOU and MSE-loss metrics. The proposed deep learning- based model surpasses the baseline model and adjacent state-of-the-art techniques with a 2.3% hike in pixel accuracy and a 1.4% increase in mean accuracy.

Keywords Human body part semantic segmentation · CDGNet · LIP · Human parsing

A. Verma (✉) · V. Tiwari · M. Lovanshi
International Institute of Information Technology (IIIT), Naya Raipur, India
e-mail: aditi21300@iiitnr.edu.in

V. Tiwari
e-mail: vivek@iiitnr.edu.in

M. Lovanshi
e-mail: mayank@iiitnr.edu.in

R. Shrivastava
Sagar Institute of Science Technology & Research, Bhopal, India
e-mail: rahul.vidisaa@gmail.com

B. Tiwari
MIT World Peace University, Pune, India

V. S. Rathore et al. (eds.), *Emerging Trends in Expert Applications and Security*,
Lecture Notes in Networks and Systems 681,
https://doi.org/10.1007/978-981-99-1909-3_19

1 Introduction

Human body part semantic segmentation is also known as human parsing. Its aim is to perform the segmentation of human body parts into fine-grained components like the right leg, left leg, hat, face, hair, scarf, upper clothes, and many more, as depicted in Fig. 1. Due to factors like intricate patterns and textures of clothing, changeable positions of people, and the scale variation of various semantic pieces, human part semantic segmentation falls under scene parsing, where pixel categorization is carried out for particular images. Potential applications such as image editing, activity recognition, emotion recognition, person reidentification, human behaviour analysis, and advanced computer vision applications are necessary to understand humans' complex semantic structure elements [1]. Human parsing is close to semantic segmentation in predicting the labels of pixels in the image. Earlier image parsing studies were primarily concerned with the spatial scale of settings. However, due to the structural limitations of convolutional layers, the geographical surroundings only provide a limited amount of contextual information. Some approaches, such as ACFNet [2], organize pixels into regions along with adding representations of the region (region representation) to the pixel representations and accomplish the best results in competitive performance comprising a variety of demanding benchmarks associated with semantic segmentation. These strategies, however, did not intentionally examine each category's spatial distribution, limiting their capacity to apprehend the distribution of separate classes and, therefore, leading to the inoperability of the rules of distribution to help parse.

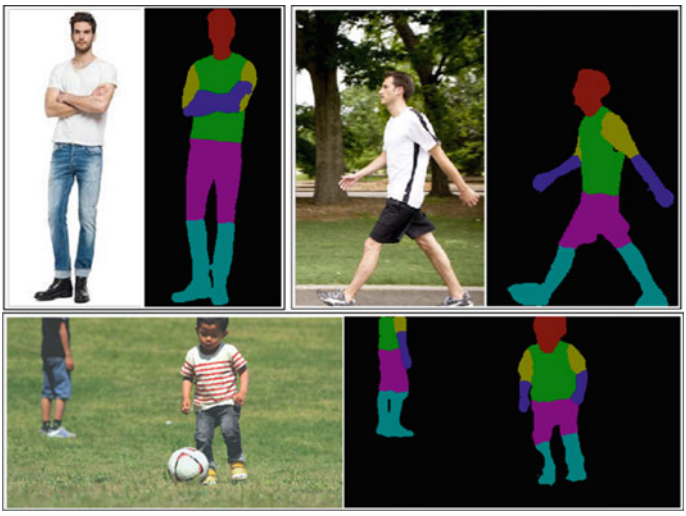

Fig. 1 Sample images with human part semantic segmentation

Nowadays, advances in FCNNs have developed various successful frameworks for human parsing tasks. Regardless, many tasks like semantic segmentation, object detection, and including a CNN-based approach require the accessibility of thoroughly annotated images for training purposes. Also, a large data set is necessary to train the framework for human parsing properly. When creating an image data set, challenges like occlusion, low quality, varying size of the image and overlapping can lead to poor prediction. Human-centric vision [3–5], human–robot interaction [6], and fashion analysis rely on the human pixel-level semantic segmentation of the human body, which is an essential task in human comprehension. In contrast to previous studies, only a few human parsing methods are generated, especially when an instance-aware environment is concerned. For instance-aware human parsing, the two paradigms currently employed are bottom-up and top-down approaches. Talking about top-down approaches, they frequently begin by locating human instance [7] ideas, which are subsequently fine-grained and processed. Bottom-up human parsers, on the other hand, are driven by the instance segmentation techniques of the bottom-up instances, followed by performing the pixel-wise grouping and instance-agnostic parsing simultaneously. Their grouping strategies span from proposal-free instance localization to graph-based super pixel association to instance edge-aware clustering.

The proposed work introduces a class distribution method called Custom-CDGNet by simplifying the complex human structure into 2D space and then converting it into vertical and horizontal 1D positional distribution according to their corresponding classes. Further, these classes teach human positional knowledge according to categories. We build a new supervision signal by collecting the vertical and horizontal-wise binarized maps from vertical and horizontal class distributions. By gathering the actual human part semantic segmentation labels in vertical and horizontal orientations and using them for supervision, we create methods, for instance, class distribution. The network takes advantage of these classes and their underlying position distribution. After that, the backbone network is then covered by a spatial guidance map that is created by combining these guided features. Our major contributions include: (1) Introduction of the new method named as Custom-CDGNet that simplifies the complexity of human parsing methods. (2) Generating class distribution labels that utilize inherent position distribution. (3) Quantitative analysis of the proposed work with different benchmark methods.

In this paper, Sect. 2 discusses the benchmarks existing in the field of semantic segmentation and human part semantic segmentation. Section 3 contains a detailed explanation of the methodology with a detailed description of the architecture, diagram and its working. Section 4 contains experimental analysis followed by a conclusion in Sect. 5.

2 Related Work

Human part semantic segmentation has recently attracted tremendous interest, with remarkable progress with advanced deep convolutional neural networks [8] and large-scale data sets. This section of the paper overviews the related works that have shown promising results. Semantic segmentation [9] deals with clustering parts of an image that belongs to the same class of an object. It is a pixel-level prediction where each pixel in an image is classified according to the category. Cityscapes [10], PASCAL VOC [10], ADE20K [11], and U-Net [12] are some of the benchmark methods for this task. In human part semantic segmentation, all the pixels in the human image are labelled accordingly. The technique used for the human part semantic segment is similar to that of scene parsing.

In human part semantic segmentation, many deep learning-based networks have established remarkable benchmarks. Ruan et al. [13] constructed CE2P [13] network with Resnet101 as the building block built, extensively used edge detail, global spatial context information, and feature resolution and obtained the best result in the LIP challenge, 2019. Yuan et al. [14] devised a representation strategy which was object-contextual for semantic segmentation and attained better performance on the LIP data set, stating that the class/label of the distinct pixel is the cluster of the category of the object to which the pixel belongs. Some studies have synthesized the prior human knowledge for human part semantic segmentation. Wang et al. [15] worked on the hierarchy of the human structure. They assembled the hierarchy for effective human part semantic segmentation. Also, Ji et al. [16] used the inherent physiological constitution of the human body by constructing a new semantic neural network tree to work on semantic segmentation of the human body parts. Zhang et al. [17] achieved excellent results by applying grammar rules in parallel and a cascaded way, using the human body's natural hierarchical structure and the relationship between various organs. Zhang et al. [18] merged keypoint positions with human semantic boundaries to enhance part semantic segmentation. These techniques rely on the specific human stance or the previous human hierarchical structure, making it challenging to guarantee comprehensiveness when several people are present or when unanticipated occlusions cover certain human body parts. A different approach, parsing R-CNN [19] introduced by Yang et al. and RP R-CNN [20] represented multi-scale features along with semantic information and attention mechanism to improve the human visual understanding of CNN to outperform in multi-human parsing and dense pose estimation [21].

3 Methodology

This section proposes a method that converts the intricately depicted human structural information in 2D space to vertical and horizontal one-dimensional (1D) positional information with their specific matching classes. The proposed approach is motivated by CDGNet [22] and attention-based models like HANet [23], CBAM [24] and SENet [25]. CBAM and SENet apprehend the extensive context of the entire image, whereas HANet considers an attention map which is height driven. The main focus of HANet is to parse urban scene images.

The proposed method is extended to classes and directional positions that expand the actual human paring labels to other guiding signals which are linked to classes and one directional location. These generated signals are crucial in leading the network for the efficient location of human body parts. It should be understood that the attention mechanism is not being utilized by the guiding signals to improve the features' presentation. Because human bodies are hierarchically built, the unique components of the human body have varied a range of apparent distributions in both horizontal and vertical directions. This paper proposes a distribution-directed network, distributed class wise that predicts the distribution of classes in dimensions categorized as vertical and horizontal while also being guided by distribution loss. The human image's projected distribution characteristic is then completely utilized to improve the feature representation for human modelling.

3.1 Class Distributions

In human body part segmentation, images with each pixel as human part labels are used in the form of training data. We compute the per-class position of distributions from individual images into the horizontal and vertical directions based on this research's labels referred to as the horizontal and vertical class distributions. The class distributions guide the network in understanding the context distribution of each category, allowing the network to evaluate various categories of spatial distributions under the constraint of the proposed distribution loss.

3.2 Distribution Network

The proposed distribution method produces each class distribution that exhibits the location of the class instances. This distribution further guides the feature representation, ultimately leading to human parsing. Taking an image frame and its features as input, we get X_i of $W \times H \times C$ where C refers to channel size. Individually squeezing the input feature X_i in the vertical and horizontal directions to extract the directional positional properties. After extracting directional characteristics, average pooling is

applied in orthogonal directions to generate labels. The proposed model uses two 1D convolution networks. The first convolution network uses 3 size kernel with its channel number to build the horizontal and vertical class distribution features. These features are then directed by making use of the new labels for the class distributions in both horizontal and vertical directions, along with the appropriate losses. At the same time, the second one-dimensional convolution network uses seven-sized kernels and its channel as input features that generate each horizontal and vertical channel distribution feature individually. Salient distribution maps are created by triggering two convolutions with sigmoid functions rather than using the softmax function. These operations, which are in the form of sequences, generate guided features on any one of the axes, either horizontal or vertical. It is denoted by:

$$A\prime h = Iu\prime p(Ah) = Iu\prime p(\sigma(conv7 \times (\delta(conv3 \times (Zh))))) \quad (1)$$

$$A\prime v = I''up(Av) = I''up(\sigma(conv7 \times (\delta(conv3 \times (Zv))))) \quad (2)$$

where δ = ReLU function, σ = sigmoid function, Z_h = horizontal class, Z_v = vertical class, A'_h = horizontal guided feature, A'_v = vertical guided feature, I'_{up}, I''_{up} = bilinear interpolation operation.

Figure 2 proposed the comprehensive architecture of the proposed work that contains the class distribution guidance (CDG) and spatial pooling (SP) module. CDG module helps in to extract horizontal and vertical classes. While SP module helps to remove the fixed-size constraint of the network. The proposed model has 3 modules, i.e. edge module, backbone, and high-resolution model. Edge module helps to extract edge between the human part by using feature fusion techniques. The backbone module focuses on the extraction of the features. While high-resolution module are used to get horizontal and vertical labels, that is feed in CDG model and SP module.

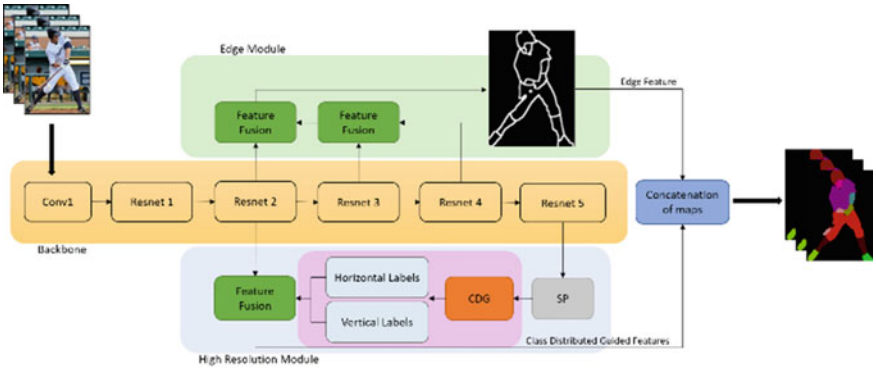

Fig. 2 Overview architecture: The figure represents the overview architecture of the proposed network. CDG stands for class distribution guidance module; SP represents pyramid spatial pooling

3.3 CDGNet Objective

The proposed methodology inspired by CDGNet and CE2P model and it outperforms to its baseline model. Objective of the proposed methodology can be considered as parsing results and edge prediction. Edge prediction is interpreted as loss of weighted cross-entropy between projected maps of the edge and its labels which is supplied by the edge module, whereas parsing results are the cross-entropy loss generated from high-resolution module parsing map and parsing labels.

4 Experimental Analysis

To demonstrate a positive performance of the suggested model in this research paper, the evaluation was carried out on the metrics like pixel accuracy, mean IoU, and overall mean accuracy. Also, class-wise mean IoU was evaluated for each of the 19 semantically labelled classes of the LIP [26], parsing data set.

4.1 Data Set Used

LIP [26], Look Into Person, is an extensive data set which focuses on the task of human parsing, also known as human part semantic segmentation. This data set contains a total of about 50,462 images which is bifurcated into 30,462 training, 10,000 testing and 10,000 validation images. The data set comprises detailed pixel-wise annotations with 19 labels describing human parts. 16 key points of 2D human poses are also included in the data set for pose estimation.

4.2 Evaluation Parameter

The proposed model evaluate results on some of the evaluation matrices, i.e. discussed below:

Mean IOU Intersection over union can be understood as the area generated by overlapping predicted segmentation result and ground truth, further dividing it by the area generated by the union of predicted segmentation result and ground. Mean IoU is usually calculated for segmenting two classes, also known as binary classes and multi-class.

Pixel accuracy (PA) is an evaluation metric representing the percentage of correctly categorized pixels in an image. This metric is used for semantic segmentation to calculate the ratio of appropriately identified pixels with the total amount of pixels present in the respective image.

Mean Accuracy is referred to as the correct predictions, which is then divided by the total input samples. While mean accuracy refers to the average accuracy of multiple classes.

4.3 Quantitative Analysis

In order to attain the highest performance in the human parsing, we performed quantitative experiments on the LIP data set in comparison with well-known human parsing algorithms. Table 1 represents the class-wise results of mean IOU on LIP data sets. In the experiment, the proposed model gives a performance of 60.52% mean IOU and outperforms existing benchmarks when compared with existing state-of-the-art frameworks. Table 2 represents the evaluated results as mean accuracy, pixel accuracy, and mean IOU comparison to other existing methods. So the proposed model claims to outperform pixel accuracy at 89.22%, mean accuracy at 72.07%, and mean IOU at 60.52 compared to other state- of-the-artwork.

5 Conclusion

This paper proposes a human part semantic segmentation method called Custom-CDGNet,

which works to attain effective and efficient semantic segmentation of human parts. This method exploits pixel labelling of each class to generate a vertical and horizontal class distribution [26] of all human parts. The knowledge of the class distribution of each class in both horizontal and vertical directions significantly benefited the learning of each pixel from the only person in the image to multiple persons as well. Performing comprehensive qualitative and quantitative analysis of Custom-CDGNet, it was found that C-CDGNet surpasses the existing state-of-the-art human parsing approaches. Also, when working on the large data set, LIP, C-CDGNet gives 89.22% pixel accuracy, 72.07% mean accuracy, and 60.52% mean IoU.

Table 1 Class-wise quantitative comparison of mean IoU with benchmark methods that performed evaluation on validation set of LIP data set. Here, each class represents the mIoU value of each human part predicted by our proposed method C-CDGNet

Method	hat	hair	glove	glass	u-cloth	dress	coat	sock	pants	j-suits	scarf	skirt	face	l-arm	r-arm	l-leg	r-leg	l-shoe	r-shoe	bkg	Avg
MMAN[27]	57.66	65.63	30.07	20.02	64.15	28.39	51.98	41.46	71.03	23.61	9.65	23.20	69.54	55.30	58.13	51.90	52.17	38.58	39.05	84.75	46.81
DeepLab[28]	56.48	65.33	29.98	19.67	62.44	30.33	51.03	40.51	69.00	22.38	11.29	20.56	70.11	49.25	52.88	42.37	35.78	33.81	32.89	84.53	44.03
Attention[29]	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
JPPNet[26]	63.55	70.20	36.16	23.48	68.15	31.42	55.65	44.56	72.19	28.39	18.76	25.14	73.36	61.97	63.88	58.21	57.99	44.02	44.09	86.26	51.37
CE2P[13]	65.29	72.54	39.09	32.73	69.46	32.52	56.28	49.67	74.11	27.23	14.19	22.51	75.50	65.14	66.59	60.10	58.59	46.63	46.12	87.67	53.10
SNT[16]	66.90	72.20	42.70	32.30	70.10	33.80	57.50	48.90	75.20	32.50	19.40	27.40	74.90	65.80	68.10	60.03	59.80	47.60	48.10	88.20	54.70
CorrPM[18]	66.20	71.56	41.06	31.09	70.20	37.74	57.95	48.40	75.19	32.37	23.79	29.23	74.36	66.53	68.61	62.80	62.81	49.03	49.82	87.77	55.33
SCHP[30]	69.96	73.55	50.46	40.72	69.93	39.02	57.45	54.27	76.01	32.88	26.29	31.68	76.19	68.65	70.92	62.28	66.56	55.76	56.50	88.36	58.62
CDGNet[22]	71.06	74.61	50.13	42.09	71.58	40.00	58.73	55.25	77.92	34.32	30.05	32.97	77.12	71.25	73.35	70.54	69.26	58.24	58.75	88.86	60.30
Our	**71.87**	**74.79**	**51.03**	**42.25**	**72.70**	**40.50**	**58.94**	54.42	**78.25**	**35.09**	29.98	**33.15**	**78.14**	**71.43**	72.44	**70.96**	**69.50**	57.47	58.24	**89.42**	**60.52**

Table 2 Comparative result of our proposed model on the LIP data set

Method	Backbone	Pixel Acc	Mean Acc	Mean IoU
CE2P[13]	Resnet 101	87.37	63.20	53.10
SNT[16]	Resnet 101	88.05	66.42	54.73
CorrPM[18]	Resnet 101	87.68	67.21	55.33
PCNet[31]	Resnet 101	–	–	57.03
HHP[15]	DeeplabV3	89.05	70.58	59.25
SCHP[30]	Resnet 101	–	–	59.36
CDGNet[22]	Resnet 101	88.86	71.49	60.30
Our	Resnet 101	**89.22**	**72.07**	**60.52**

References

1. Rochan M (2018) Future semantic segmentation with convolutional lstm. arXiv:1807.07946
2. Zhang Fan et al (2019) Acfnet: attentional class feature network for semantic segmentation. In: Proceedings of the IEEE/CVF international conference on computer vision, pp 6798–6807
3. Ebadi SE et al (2021) PeopleSansPeople: a synthetic data generator for human-centric computer vision. arXiv:2112.09290
4. Arshad A, Tiwari V, Lovanshi M, Shrivastava, R (2023) Role identification from human activity videos using recurrent neural networks. In: Proceedings of 8th IEEE international women in engineering (WIE) conference on electrical and computer engineering (WIECON-ECE)
5. Lovanshi M, Tiwari V (2023) Human pose estimation: benchmarking deep learning-based methods. In: Proceedings of the IEEE conference on interdisciplinary approaches in technology and management for social innovation
6. Hu H, Jaime FF (2022) Active uncertainty learning for human-robot interaction: an implicit dual control approach. arXiv:2202.07720
7. Shrivastava R, Tiwari V, Jain S, Tiwari B, Kushwaha AKS, Singh VPA (2022) role-entity based human activity recognition using inter-body features and temporal sequence memory. IET Image Process
8. Choudhary M, Vivek T, Venkanna U (2020) Enhancing human iris recognition performance in unconstrained environment using ensemble of convolutional and residual deep neural network models. Soft Comput 24(15):11477–11491
9. Bose K, Shubham K, Tiwari V, Patel KS (2023) Insect image semantic segmentation and identification using UNET and DeepLab V3+. In: ICT infrastructure and computing. Springer, Singapore, pp 703–711
10. Wang P, Chen P, Yuan Y, Liu D, Huang Z, Hou X, Cottrell G (2018) Under- standing convolution for semantic segmentation. In: 2018 IEEE winter conference on applications of computer vision (WACV), pp 1451–1460
11. Zhou B, Zhao H, Puig X, Xiao T, Fidler S, Barriuso A, Torralba A (2019) Semantic understanding of scenes through the ade20k dataset. Int J Comput Vision 127(3):302–321
12. Ronneberger O, Fischer P, Brox T (2015) U-net: convolutional networks for biomedical image segmentation. In: International conference on medical image computing and computer-assisted intervention, Springer, Berlin, pp 234–241
13. Ruan T, Liu T, Huang Z, Wei Y, Wei S, Zhao Y (2019) Devil in the details: Towards accurate single and multiple human parsing. In: Proceedings of the AAAI conference on artificial intelligence, vol 33, pp 4814–4821
14. Yuan Y, Huang L, Guo J, Zhang C, Chen X, Wang J (2018) Ocnet: object context network for scene parsing. arXiv:1809.00916

15. Wang W, Zhu H, Dai J, Pang Y, Shen J, Shao L (2020) Hierarchical human parsing with typed part-relation reasoning. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 8929–8939
16. Ji R, Du D, Zhang L, Wen L, Wu Y, Zhao C, Huang F, Lyu S (2020) Learning semantic neural tree for human parsing. In: European conference on computer vision. Springer, Berlin, pp 205–221
17. Zhang X, Chen Y, Zhu B, Wang J, Tang M (2020) Blended grammar network for human parsing. In: European conference on computer vision. Springer, Berlin, pp 189–205
18. Zhang Z, Su C, Zheng L, Xie X (2020) Correlating edge, pose with parsing. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 8900–8909
19. Yang L et al (2019) Parsing R-CNN for instance-level human analysis. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 364–373
20. Yang Lu et al (2020) Renovating parsing R-CNN for accurate multiple human parsing. In: European conference on computer vision. Springer, Cham, pp 421–437
21. Güler RA, Natalia N, Iasonas K (2018) Densepose: dense human pose estimation in the wild. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 7297–7306
22. Liu K, Choi O, Wang J, Hwang W (2022) Cdgnet: class distribution guided network for human parsing. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 4473–4482
23. Choi S, Kim JT, Choo J (2020) Cars can't fly up in the sky: Improving urban-scene segmentation via height-driven attention networks. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 9373–9383
24. Woo S, Park J, Lee J-Y, Kweon IS (2018) Cbam: convolutional block attention module. In: Proceedings of the European conference on computer vision (ECCV), pp 3–19
25. Hu J, Shen L, Sun G (2018) Squeeze-and-excitation networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 7132–7141
26. Liang X, Gong K, Shen X, Lin L (2018) Look into person: joint body parsing & pose estimation network and a new benchmark. IEEE Trans Pattern Anal Mach Intell 41(4):871–885
27. Luo Y, Zheng Z, Zheng L, Guan T, Yu J, Yang Y (2018) Macro-micro adversarial network for human parsing. In: Proceedings of the European conference on computer vision (ECCV), pp 418–434
28. Chen L-C, Papandreou G, Kokkinos I, Murphy K, Yuille AL (2017) Deeplab: semantic image segmentation with deep convolutional netsatrous convolution, and fully connected crfs. IEEE Trans Pattern Anal Mach Intell 40(4):834–848
29. Chen L-C, Yang Y, Wang J, Xu W, Yuille AL (2016) Attention to scale: Scaleaware semantic image segmentation. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 3640–3649
30. Li P, Xu Y, Wei Y, Yang Y (2020) Self-correction for human parsing. IEEE Trans Pattern Anal Mach Intell
31. Zhang X, Chen Y, Zhu B, Wang J, Tang M (2020) Part-aware context network for human parsing. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 8971–8980