CDGNet: Class Distribution Guided Network for Human Parsing

Kunliang Liu[1,2], Ouk Choi[3], Jianming Wang[2*], and Wonjun Hwang[1*]

[1]Ajou University, Korea, [2]Tiangong University, China, [3]Incheon National University, Korea

tjpulkl@ajou.ac.kr, ouk.choi@inu.ac.kr, wangjianming@tiangong.edu.cn, wjhwang@ajou.ac.kr

Abstract

The objective of human parsing is to partition a human in an image into constituent parts. This task involves labeling each pixel of the human image according to the classes. Since the human body comprises hierarchically structured parts, each body part of an image can have its sole position distribution characteristic. Probably, a human head is less likely to be under the feet, and arms are more likely to be near the torso. Inspired by this observation, we make instance class distributions by accumulating the original human parsing label in the horizontal and vertical directions, which can be utilized as supervision signals. Using these horizontal and vertical class distribution labels, the network is guided to exploit the intrinsic position distribution of each class. We combine two guided features to form a spatial guidance map, which is then superimposed onto the baseline network by multiplication and concatenation to distinguish the human parts precisely. We conducted extensive experiments to demonstrate the effectiveness and superiority of our method on three well-known benchmarks: LIP, ATR, and CIHP databases. [†]

1. Introduction

Human parsing involves segmenting human bodies into constituent parts, such as the head, body parts, and clothing items. It belongs to the field of scene parsing, where per-pixel categorization is performed for a given image. Human parsing is highly challenging owing to the complex textures and styles of clothes, deformable human poses, and scale diversity of different semantic parts. Human parsing enables fine-grained semantic segmentation, benefits human understanding, and supports human-centric applications.

Since the emergence of fully convolutional networks (FCNs) [26], various studies have developed several solutions from different perspectives to boost the performance of this dense prediction task. Several previous meth-
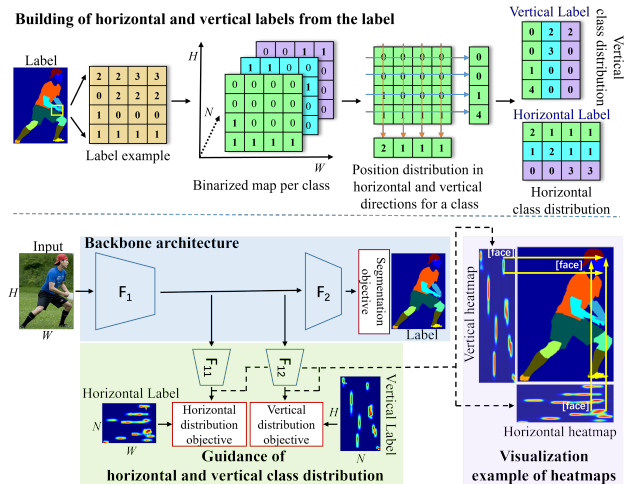

Figure 1. **Guidance of class distribution**. (Top) We build the novel horizontal and vertical class distribution labels from the original label to simplify the 2D spatial human parsing problem with 1D positional labels in the horizontal and vertical directions. (Bottom-left) We directly teach the backbone network with our horizontal and vertical class distribution labels as a positional guidance signal of human parts. (Bottom-right) For instance, the horizontal and vertical face positions are correctly estimated as observed in the visualization heatmap examples.

ods [4,8,9,14,18,21,30,36–38,44] using both spatial and semantic contexts have achieved great successes in scene segmentation; however, human parsing has unique characteristics that make it much harder than the conventional pixel-level segmentation tasks. Specifically, the human body comprises symmetrical and structural parts. The left and right arms, and the left and right shoes, have a similar appearance as well as a similar vertical position but must be separated into different classes. Naturally, in this case, we utilize the horizontal position difference between the left and right human parts. Furthermore, the scarf and glasses are too small parts to be segmented; however, most of them are on the face or the neck, not on the legs, and we can obtain better performance when using this structural knowledge successfully. Several works took advantage of human characteristics for achieving better performance in human

parsing. Ji et al. [19] designed a novel semantic neural tree to encode the physiological structure of the human body and achieved competitive results. Wang et al. [31, 32] exploited deep graph networks and hierarchical human structures to capture the relation information of human parts and obtained better performances. These mechanisms involve designing a complex semantic tree or message-passing network that leads to heavy computing complexity while improving performance. Moreover, building different tree structures for diverse datasets limits the deployment of these methods. Zhang et al. [43] utilized human pose and non-local mechanism to achieve good performance, but it required accurate human pose information in advance. Consequently, most recent works have attempted to solve complex human parsing problems by utilizing complex modules such as graph network [32] and human pose estimator [43].

In this paper, we attempt to solve the human parsing problem following the *divide and conquer* strategy. We simplify the human structure information complexly represented in 2D space into horizontal and vertical 1D position information with the corresponding classes. To this end we propose a novel class distribution guided network (CDGNet) taught by human categorical positional knowledge represented in the horizontal and vertical directions. As shown at the top of Fig. 1, we first build the horizontal and vertical class distributions as new supervision signals by accumulating the horizontal-wise and vertical-wise binarized map for each class from the original label. We squeeze the baseline feature in the horizontal and vertical directions and then teach them using the corresponding class distribution labels, as described in the bottom of Fig. 1. For example, the face candidate region is successfully represented by the face class distributions at the right bottom of Fig. 1. Note that we can use all the classes without selecting a few specific classes for learning the model because we have simplified the complex human parsing problem. Finally, in order to accomplish the precise human paring results, we merge these guided features and then superimpose them on the backbone network features.

The major contributions of our work are summarized as follows:

• We simplify the complex spatial human parsing problem into the horizontal and vertical positions of human parts individually. Accordingly, we build the class distribution labels in the horizontal and vertical directions as new supervision signals from the original label of human parsing.

• Using these class distribution labels, we propose the CDGNet that guides the backbone network toward exploiting the intrinsic position distribution of human parts.

• We verify the significant performance improvement gained by the proposed method through extensive experiments on LIP, ATR, and CIHP benchmarks.

2. Related Works

Semantic Segmentation: Human parsing is a fine-grained semantic segmentation task in which all pixels of the human image are labeled. The method utilized in human parsing is similar to that used for semantic segmentation, which predicts each pixel labeling in the scene. Earlier scene parsing studies mainly focused on the spatial scale of contexts [1–6, 8, 9, 14, 16–18, 20, 29, 33, 34, 40, 43, 44, 46]. However, owing to the limitations of the structure of convolutional layers, the spatial context provides insufficient contextual information. Methods such as OCR [37] and ACFNet [39] group the pixels into a set of regions, augment the pixel representations by region representations to which the pixel belongs, and gain competitive performance on various challenging semantic segmentation benchmarks; however, these methods did not intentionally consider the spatial distribution of each category, confining the ability to capture the distribution of different classes, and cannot utilize the distribution rule to benefit parsing efficiently.

Human Parsing: Several deep learning-based methods have achieved significant improvements in human parsing. Ruan et al. [30] designed a CE2P network based on ResNet-101 that fully leveraged feature resolution, global spatial context information, and edge detail. It achieved its first place in LIP Challenge in 2019. Given the label of a pixel is the category of the object to which the pixel belongs, Yuan et al. [38] implemented an object-contextual representation approach for semantic segmentation and achieved an even higher performance on LIP. Other studies have combined additional human prior information for human parsing. For instance, Wang et al. [32] assembled the compositional hierarchy of human bodies for efficient and complete human parsing, and Ji et al. [19] exploited the intrinsic physiological structure of the human body by designing a novel semantic neural tree for human parsing. Utilizing grammar rules in a cascaded and parallel manner, Zhang et al. [42] employed the inherent hierarchical structure of the human body and the relationship of different human parts to achieve impressive human parsing results. Zhang et al. [43] combined human semantic boundaries and key-point locations to improve human parsing. These methods either depend on the prior human hierarchical structure, or the precise human pose, which makes it difficult to guarantee generality if there are multiple persons or if human parts are covered by unexpected occlusions.

3. Class Distribution Approach

In this section, we propose an efficient method that utilizes human part class distributions for human parsing without additional heavy networks, e.g., human pose detection [43] and complex assumption [32]. Our method is catalyzed by attention-based methods such as SENet [15],
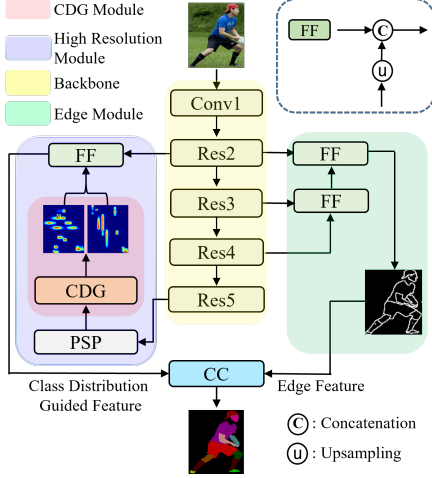
Figure 2. **Overview of our network. FF**: feature fusing. **PSP**: pyramid spatial pooling. **CC**: concatenation and convolution of feature map. **CDG**: our class distribution guidance module.

CBAM [35] and HANet [7]. SENet and CBAM captured the global context of the entire image for general purposes. HANet considered solely a height-driven attention map because it focused on urban scene images.

In this paper, rather than employing the attention methods, we extend the original human parsing label to another supervision signals related to classes and one directional positions. These generated signals play a pivotal role in guiding the network to locate the human parts efficiently. Note that the attention mechanism has not used the supervision signals for enhancing the feature representation. Because human bodies are hierarchically structured as described in [32], which means that it has significantly strict spatial distribution features, the distinct parts of a human body have different manifest distributions in the vertical and horizontal dimensions. We propose a class distribution guided network that predicts the class distribution in the vertical and horizontal dimensions under the supervision of class distribution loss. The predicted distribution characteristic of the human image is then fully employed to enhance the feature representation for human modeling. The overall proposed procedure is summarized in Fig. 2.

3.1. Horizontal and Vertical Class Distribution

In human parsing, images with per-pixel human-part labels are given as a training dataset. From the labels, we calculate the per-class positional distributions within each image in the horizontal and vertical directions, respectively, which are called horizontal and vertical class distributions in this paper. The class distributions guide the network to learn the distribution context of each category, which forces the network to consider different category spatial distributions under the constraint of the proposed distribution loss. Given an input image $\boldsymbol{I}$ and the corresponding label $\boldsymbol{G}$, we

encode each class integer label to one-hot encoding, generating a matrix $\boldsymbol{M}$ of $H \times W \times N$ size, where H and W are the input image height and width, respectively, and N is the class number. As the top part of Fig. 1 indicates, in each channel, the number of '**1**' is counted along the horizontal and vertical directions, generating the vertical and horizontal class distributions. Therefore, the two distributions, $\boldsymbol{G_{D_h}}$ of size $W \times N$ and $\boldsymbol{G_{D_v}}$ of size $H \times N$ can be obtained by the following equations;

$$\boldsymbol{G_{D_h}}(w,n) = \sum_{i=1}^{H} \boldsymbol{M}_{i,w,n}, \ \boldsymbol{G_{D_v}}(h,n) = \sum_{i=1}^{W} \boldsymbol{M}_{h,i,n},$$
(1)

where $h \in \{1 \sim H\}, n \in \{1 \sim N\}, w \in \{1 \sim W\}, \boldsymbol{G_{D_h}}$ and $\boldsymbol{G_{D_v}}$ are the horizontal and vertical class distributions as new supervision signals for human parsing, respectively.

3.2. Class Distribution Guided Network

Our CDGNet produces each class distribution that reveals where and to what extent the class instance exists and correctly guides the feature representation for human parsing. The structure of CDGNet is illustrated in Fig. 3. When inputting a feature, X_i of $W \times H \times C$ where C is the channel size, we squeeze the input feature in the horizontal and vertical directions, individually, to extract the directional position characteristics. Accordingly, we employ average pooling in orthogonal directions, e.g., vertical average pooling for the horizontal feature $\boldsymbol{Z}_h$ and horizontal average pooling for the vertical feature $\boldsymbol{Z}_v$. After average pooling, one 1D convolution with kernel size 3 followed by Batch Normalization (BN) layer is leveraged to decrease the channel number, i.e., $C/2$. Because the number of classes is generally smaller than the channel size, it is necessary to reduce the channel size by half as a buffer.

To generate the horizontal and vertical class distribution features, we utilize 1D convolution with kernel size 3 with its channel number, N, and we guide these features using the new labels regarding the horizontal and vertical class distributions through the corresponding losses. Meanwhile, another 1D convolution with kernel size 7 and the same channel number with the input feature, C, is applied to produce each channel distribution feature in the horizontal and vertical directions, separately. The two convolutions are activated by the $sigmoid$ function, instead of the $softmax$ function to produce the salient distribution maps. The reason is that multiple classes exist in each row of $\boldsymbol{Z}_h$ and $\boldsymbol{Z}_v$. These series of operations for generating the guided features on the horizontal or vertical dimensions of each channel can be denoted as:

$$\boldsymbol{A}_h^{'} = I_{up}^{'}(\boldsymbol{A}_h) = I_{up}^{'}(\sigma(conv_{7\times}(\delta(conv_{3\times}(\boldsymbol{Z}_h))))),$$
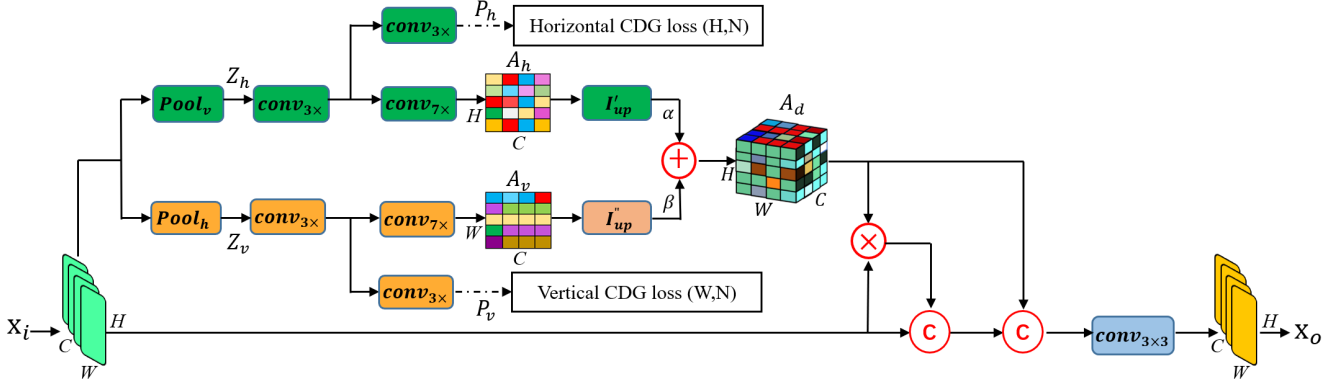(2)

Figure 3. **Architecture of the proposed CDGNet.** X_i: input feature, Z_h and Z_v: horizontal and vertical features, A_h and A_v: horizontal and vertical guided features, A_d: spatial guidance map, X_o: CDG feature, $Pool_v$: pooling along vertical direction to obtain horizontal distribution, $Pool_h$: pooling along horizontal direction to obtain vertical distribution, $conv_{3\times3}$: 3×3 kernel-based 2D convolution, $conv_{3\times}$, and $conv_{7\times}$: 1D convolution with kernel size 3 and 7, I'_{up} and I''_{up}: upsampling horizontal and vertical guided features, N: the category number of human part, H, W and C: height, width and channel number of the input feature map, α and β: coefficients of horizontal and vertical guided features when aggregating them. $\oplus$: element-wise summation, $\otimes$: element-wise multiplication of the matrix, and Ⓒ: feature concatenation.

$$A'_v = I''_{up}(A_v) = I''_{up}(\sigma(conv_{7\times}(\delta(conv_{3\times}(Z_v))))), \tag{3}$$

where σ is a *sigmoid* function, and δ is a *ReLU* function. The guided feature A_h involves the position-related features for each class in the channel according to the horizontal direction whereas the guided feature A_v includes the vertical position-related class features in the channel because they are all guided by the corresponding CDG loss functions. Bilinear interpolation operations I'_{up} and I''_{up} are then applied to the corresponding guided features to expand their other dimensions to have the same dimension as the input feature X_i. The extended guided features A'_h and A'_v are weighted by coefficients α and β and merged by the element-wise sum operation to combine the separately learned models. We adopt the element-wise multiplication and concatenation sequentially to properly reflect the spatial guidance map A_d to the input feature. Using this scheme, the input feature X_i can be transformed into a CDG feature X_o. Finally, a 3×3 kernel-based convolution is adopted to fit the output feature size identical to the input feature size. The formula is as follows:

$$A_d = \alpha \times A'_h + \beta \times A'_v, \tag{4}$$

where α and β are the learnable weight parameters.

3.3. Class Distribution Guided Loss

The human body is non-rigid, and a few human parts can be occluded during vigorous exercise. These facts pose a difficult challenge in improving the accuracy of the human parsing model utilizing only 2D human body labels. The

edge generated from the label [30, 43] is used to clarify the boundaries between human parts to solve this problem. If we first know where it is most likely to determine the different human part classes in the feature map, the problem can be alleviated. In this respect, we introduce the CDG loss for teaching the network to model each category distribution, where pixels on each row in vertical and horizontal distribution maps indicate the related category distribution in both directions.

The predicted class distribution maps are P_h of size W and N, and P_v of size H and N as shown in Fig. 3. Assume that each position in the predicted horizontal class distribution map is $\{p_{i,j} \in P_h, i \in [1, W], j \in [1, N]\}$, and in the vertical class distribution map is $\{p'_{i,j} \in P_v, i \in [1, N], j \in [1, H]\}$. Accordingly, assuming that the corresponding positions of the horizontal class distribution label are $\{g_{i,j} \in G_{D_h}, i \in [1, W], j \in [1, N]\}$, and the vertical class distribution label is $\{g'_{i,j} \in G_{D_v}, i \in [1, N], j \in [1, H]\}$. The L2 loss between the prediction and label can be denoted as:

$$l_h = \frac{1}{W \times N} \sum_{i=1}^{W} \sum_{j=1}^{N} (p_{i,j} - g_{i,j})^2, \tag{5}$$

$$l_v = \frac{1}{N \times H} \sum_{i=1}^{N} \sum_{j=1}^{H} (p'_{i,j} - g'_{i,j})^2. \tag{6}$$

The complete CDG loss l_{CDG} is denoted as follows;

$$l_{CDG} = \theta \times l_h + \phi \times l_v, \tag{7}$$

where θ and ϕ are the balance weights for the horizontal and vertical CDG losses, respectively. The weights are set as $\theta = 1$ and $\phi = 1$ for the simplicity in this paper.

3.4. Training Objectives

In this paper, CE2P [30] was used as the baseline network. The CE2P output includes two parsing results and one edge prediction. Thus, the total loss of CE2P can be denoted as:

$$L_{baseline} = L_{parsing} + L_{edge} + L_{edge-parsing}, \quad (8)$$

where L_{edge} is the weighted cross-entropy loss between the predicted edge map generated using the edge module and edge label map. Here, $L_{parsing}$ is cross-entropy loss between the parsing map of the high-resolution module and human parsing label, and $L_{edge-parsing}$ is another cross-entropy loss between the parsing label and the final parsing result, after combining the feature maps of edge perceiving module. Our method appends after the pyramid pooling module, as shown in Fig. 2. Thus, the final loss function can be formulated using equations (7) and (8) as follows:

$$L = \tau \times L_{baseline} + \gamma \times l_{CDG}, \quad (9)$$

where τ, and γ are weight hyperparameters. The weights were set to $\tau = 1$ and $\gamma = 40$.

3.5. Discussion

Recently, attention-based semantic segmentation methods [8, 21, 33, 36, 46] have achieved successful improvements, but they require too much complexity in the computation. These methods involve leveraging matrix multiplication to capture the correlation between all pixels and classes. Specifically, assuming the feature size to be $H \times W$, the attention weight matrix of these methods should have a size of $(H \cdot W) \times (H \cdot W)$. It should be noted that the segmentation results in better performance as the resolution of the input image increases. However, our CDGNet statistically calculates each class distribution in the horizontal and vertical directions. In contrast, to specifically consider the distribution properties of each class, the required distribution feature size is solely $H \times C + C \times W$. Note that the feature size $((H + W) \times C)$ of our method is much smaller than that of the attention weight matrix. The memory requirement decreases significantly in this case. When scaling the feature based on the distribution map, solely matrix element-wise multiplication and addition are required as shown in Fig. 3, which is also more computationally efficient than the matrix multiplication used in the attention-based method. Our CDGNet is, therefore, considerably lightweight compared to attention-based approaches.

Other related works that involve semantic tree [19] or human pose [43] require constructing different trees or human-labeling additional annotations for the various datasets, which constrains the generality of deployment. By comparison, our model does not need to adjust any parts of the architecture when applied to different datasets and can

serve as a plugin module for any scene parsing network, as shown in the bottom part of Fig. 1. Because most of the computational complexity occurs in 1D convolution, it is considerably lightweight and generic. Regarding the human pose-based method, despite the additional utilization of the pose estimator, the information obtained is only about the location of the head, body, arms, and legs, which are the pivotal human parts. Moreover, the final human parsing performance depends on the performance of the pose estimator. Our algorithm utilizes the horizontal and vertical class distribution maps of each part of the human, which provides not only the statistical position of each part, but also the distribution range. Our method utilizes the distribution maps of all target classes including hat, sock, etc., which have not been included in the results of the pose estimator. The non-local module has been utilized to capture the long-range contextual appearance information for each pixel based on self-attention methods [8, 34, 43]. Our network can employ the non-local module, achieving higher performance. We conjecture that our class spatial distribution feature and the non-local feature are complementary to each other to provide better performance for human parsing.

4. Experiments

4.1. Datasets

Look Into Person (LIP): The LIP dataset [12] was utilized in the LIP Challenge 2016 for human parsing tasks. In total, 50,462 images were provided, including 30,462 images for training and 10,000 for validation. These images are finely labeled at the pixel level with 19 semantic human part classes (including 6 body parts and 13 items of clothing) and one background category.

Active Template Regression (ATR): The ATR dataset [25] contains 18 semantic category labels (including face, sunglasses, hat, scarf, hair, upper clothes, left arm, right arm, belt, pants, left leg, right leg, skirt, left shoe, right shoe, bag, dress, and background). In total, 17,700 images were included in ATR. Following [43], 16,000 images were utilized for training, 1,000 for testing, and 700 for validation.

Crowd Instance-level Human Parsing (CIHP): The CIHP dataset [11] is a large-scale dataset that provides 38,280 multi-person images with 20 semantic parts, including the background. The images in the CIHP were collected from a real-world scenario. Persons in the images appear with challenging poses and viewpoints, heavy occlusions, and show in a wide range of resolutions [10]. The dataset is elaborately annotated to benefit the semantic understanding of multiple people in a real situation. It comprises 28,280 training, 5,000 validation and 5,000 test images.

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
Attention [5]	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
DeepLab [3]	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
MMAN [28]	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
SS-NAN [45]	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
JPPNet [24]	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 [30]	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 [19]	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 [43]	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 [23]	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	67.28	66.56	55.76	56.50	88.36	58.62
Ours	_70.71_	74.43	**50.68**	_41.72_	_70.92_	38.51	57.60	_54.30_	77.65	_33.86_	**30.59**	_31.98_	76.89	_70.93_	_73.03_	70.30	68.92	58.17	_58.67_	88.83	_59.93_
Ours†	**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

Table 1. Per-class intersection over union (IoU) comparison of the validation set of LIP. Here, † is the parsing result averaged over the multiscaled image pyramids with flipping in the inference time. The result of SCHP [23] is made by a single image in the inference time. The best values are marked in bold, and the second-best values are marked with an underline.

4.2. Experimental Settings

Implementation Details: We employed the basic structure and network settings of CE2P [30] as the baseline. The feature from the backbone with 512 channels feeds into two parallel 1D spatial pooling along the height or width axes. The channel number of the horizontal and vertical features is reduced to 256 through a 1D convolution with a kernel size of 3. Subsequently, we generate the predicted class distribution map, P_h of size $H \times N$ and P_v of size $N \times W$, using the convolution with kernel size 3 and sigmoid, which is guided by the proposed CDG loss. Meanwhile, with a 1D convolution of kernel size 7 and $sigmoid$ activation, the channel numbers of the horizontal and vertical features are increased to 512. Subsequently, the guided features are upsampled using the bilinear interpolation to $H \times W \times 512$, and they are merged into the spatial guidance map by element-wise summation. We perform matrix element-wise multiplication between the spatial guidance map and the input feature. We finally reduce its channel number to 512 using $3 \times 3 \ convolution + BN + ReLU$ to extract the CDG feature.

Data Augmentation: In the training procedure, we adopt the mean subtraction, random scaling in the range of $[0.5, 1.25]$, random color jittering, and random left-right flipping as the basic data augmentation. We randomly crop the large image or pad the small images into a fixed size for training (i.e., 473×473 for LIP and CHIP, 512×512 for ATR).

Training: We adopt ResNet-101 [13] pre-trained using the ImageNet database [22] as the backbone. After appending the CDG module in the context embedding module of CE2P [30], the network is trained for 150 epochs on the LIP and CIHP datasets, and 250 epochs on the ATR dataset. Stochastic gradient descent (SGD) is utilized as the optimizer, and the initial learning rate is set to $3e^{-3}$. We adopt the polynomial learning rate strategy $r = (1 - \frac{cur\_iter}{total\_iter})^{0.9}$, where cur_iter and $total\_iter$ represent the current iteration and total iteration numbers, respectively. The momentum is set to 0.9, and the weight decay is $5e^{-4}$. We adopt cross-entropy loss when training on all datasets.

Inference: In the inference time, the pixel accuracy (pixAcc), mean accuracy, and mean pixel Intersection-over-union (mIoU) are leveraged as the evaluation metrics for the LIP dataset, pixel accuracy, precision, recall, and F-1 scores for the ATR dataset, and mIoU for the CIHP dataset. Similar to [10, 23, 32, 41], we averaged the results using 3-scale image pyramids of different scales $[0.75, 1.0, 1.25]$ and flipping for further performance improvement.

4.3. Quantitative and Qualitative Experiments

To achieve the best performance in human parsing, in this paper, we make use of mIOU loss [23, 32] and non-local module [34] with our proposed method. Quantitative experiments were performed by comparison with well-known human parsing methods using the LIP, ATR, and CIHP datasets. Through these experimental results, we verified that our method can achieve meaningful and precise performance regardless of the difficulty of the human image and number of persons.

Performance on LIP database: Table 1 presents the performance comparison of the proposed method to other methods on the LIP validation set. Basically, our method achieves the best performances compared with others as well as the baseline method through all 20 classes of LIP dataset. What we should note here is that, compared with others, we significantly improve the IoUs of small parts of human parsing, for example, glasses, scarves, and shoes. In this respect, we can infer that our positional distributions using all classes are better than the positions of the selected human parts estimated by the complex pose estimator of CorrPM [43]. From the viewpoint of the IoU of the scarf, our method achieves approximately 6% better than CorrPM [43] and 4% compared to SCHP [23]. It is mainly because the position of the scarf could be determined around the head and upper body, and we naturally have a higher chance to teach the network using the supervision signals for the horizontal and vertical class distributions to possess the corresponding information. Consequently, our CDGNet

Method	Pixel Acc.	Mean Acc.	mIoU
CE2P [30]	87.37	63.20	53.10
SNT [19]	88.05	66.42	54.73
Double Attention [8]	-	-	55.12
CorrPM [43]	87.68	67.21	55.33
BGNet [42]	-	-	56.82
ISNet [21]	-	-	56.96
MCIBISS [20]	-	-	56.99
PCNet [41]	-	-	57.03
HHP [32]	**89.05**	<u>70.58</u>	59.25
SCHP [23]	-	-	<u>59.36</u>
Ours	<u>88.86</u>	**71.49**	60.30

Table 2. Comparison of different methods on the validation set of the LIP dataset. The bold is the best one and the underline is the second best.

Method	Acc.	F.G.Acc.	Pre.	Recall	F-1
DeepLabV2 [3]	94.42	82.93	69.24	78.48	73.53
Attention [5]	95.41	85.71	81.30	73.55	77.23
CoCNN [25]	96.02	83.57	84.59	77.66	80.14
TGPNet [27]	96.45	87.91	83.36	80.22	81.76
CorrPM [43]	<u>97.12</u>	**90.40**	**89.18**	83.93	86.12
HHP [32]	96.87	89.23	86.17	**88.35**	**87.25**
Ours	**97.39**	<u>90.19</u>	<u>87.46</u>	<u>86.87</u>	<u>87.16</u>

Table 3. Comparison of accuracy, foreground accuracy, precision, recall and F-1 scores on the ATR test set. The bold font is the best value, and the underline is the second.

exhibits the state-of-the-art performance (60.30% mIoU) on LIP, and our result is approximately 1% better than the human part relation-based HHP [32] (59.25% mIoU) and the label self correction-based SCHP [23] (59.36% mIoU), as summarized in Table 2. When compared with the baseline (CE2P), our method achieves 7.20% higher in terms of mIoU and 1.5% and 8.2% higher in terms of pixel accuracy and mean accuracy, respectively. We conclude from these results that our method, which considers the class distributions correlated with the whole class, can yield more promising results than the others using a few relations of the selected human parts in human paring.

Performance on ATR database: The ATR database comprises a relatively small number of images compared with the LIP dataset, and most of the images are captured with a single person's frontal pose. As presented in Table 3, our method achieves the best performance in terms of pixel accuracy and the second rank for foreground pixel accuracy, precision, recall, and F-1 measures. Specifically, compared with CorrPM [43], our method showed better results, except for foreground pixel accuracy, and precision. However, note that the performance gap between our method and CorrPM in the LIP database is approximately 5% at mIoU. When comparing our method with HHP [32], our method achieves better results in terms of pixel accuracy, foreground accuracy, and precision, but lower results at recall and F-1. We note that our method performs 1% better on the LIP dataset.

Method	Backbone	mIoU
PGN [11]	DeepLabV2	55.80
Graphonomy [10]	DeepLabV3+	58.58
M-CE2P [30]	ResNet101	59.50
CorrPM [43]	ResNet101	60.18
SNT [19]	ResNet101	60.87
PCNet [41]	ResNet101	<u>61.05</u>
Ours	ResNet101	**65.56**

Table 4. Performance comparison with state-of-the-art method on CIHP dataset. The bold font is the best value, and the underline is the second.

Performance on CIHP database: In the multi-person-based database CIHP, our CDGNet achieves the best performance of 65.56% among the previous works based on ResNet-101, and outperforms the baseline and PCNet by 6.1% and 4.5%, respectively, as presented in Table 4. Note that improving the multiple human parsing task is highly challenging, but we do not design additional methods for multiple person-based human parsing and obtain 5.5% higher performance than CorrPM based on the pose information. Moreover, the input image size is 473×473, which is much smaller than that of PCNet, which leverages an image size of 512×512. In the CIHP, we are required to train a more general human parsing model for the case of multiple persons. For example, when multiple persons stand close together, occlusion frequently occurs, which does not occur in a single person. Our method can overcome this problem because the horizontal class distribution is statistically made by accumulating the features of multiple persons existed in an image. In this respect, the noise of one or two persons among several persons does not significantly affect the overall performance of the proposed class distributions.

4.4. Qualitative Results

In this section, we first verify that the proposed network correctly predicts the horizontal and vertical class distributions. As shown in Fig. 4, the predicted results are very similar to Ground Truth (GT) generated from the original human parsing label. Therefore, we can infer that the horizontal and vertical labels guided the proposed method well during the training procedure, and horizontal and vertical class distributions can reflect individually where each class exists with high probability. Fig. 5 validates that our method achieves improved performance on various human parts with manifest distribution characteristics, such as the right arm in the first row, right leg in the second row, and glasses in the third row. In the fourth row examples of Fig. 5, our method and CorrPM indicate the correct positions of the left and right shoes while the baseline does not; thus, we can infer that the proposed class distribution guided network can distinguish the left and right parts of human parsing successfully. In the last row example, the baseline makes other

	Method			Pixel Acc.	Mean Acc.	mIoU
B	G	I	N			
✓	-	-	-	87.37	63.20	53.10
✓	✓	-	-	88.41	68.87	57.72
✓	✓	✓	-	88.53	70.89	59.04
✓	✓	✓	✓	88.65	71.98	59.93

Table 5. Each component of our method is evaluated on the LIP validation set, including baseline (B), CDG module (G), IOU loss (I), non-local module (N).
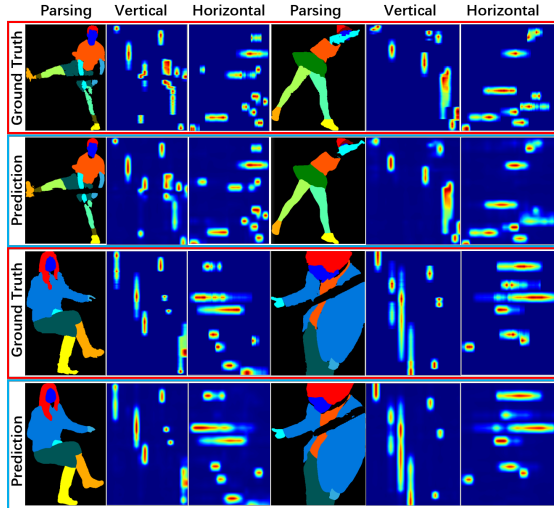

Figure 4. Visualization of the predicted class distribution map. CDGNet generates the horizontal and vertical distributions of the human parts under the supervision of corresponding distribution labels. Deeper color denotes higher probability of class existence.

errors on the left and right shoes, and CorrPM indicates the poor result owing to the occlusion problem. However, our method correctly predicts the human parsing result.

4.5. Ablation Studies

As Table 5 presents, when we soley integrate the CDG module with the baseline, it makes a dramatic improvement in terms of mIoU, from 53.1% to 57.72%. Note that this performance is significantly higher than that of attention-based and context-based methods, the results of which are presented in Table 2. For instance, it is 2.5% higher than that of the double attention method [8] and 0.72% higher than the latest image- and semantic-level context method [21]. By further adopting mIoU loss that has been commonly utilized in the latest works [23, 32] to achieve the state-of-the-art accuracy, we achieved a higher performance of 59.04%. Finally, we verify that our CDGNet can learn complementary features to the non-local module and unify the non-local module with CDGNet, which results in the performance improvement from 59.04% to 59.93% in terms of mIoU.

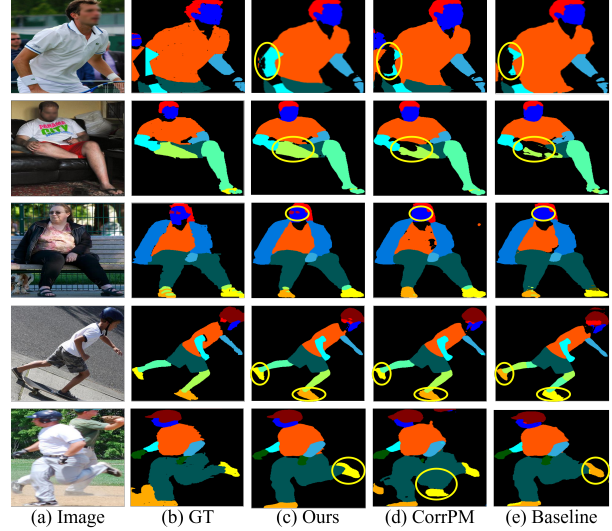(a) Image   (b) GT   (c) Ours   (d) CorrPM   (e) Baseline

Figure 5. Visualization examples of different methods on the LIP validation dataset. We point out the differences between ours and others with yellow ellipses. To avoid the negative societal impact, when showing faces in this paper, face regions are blurred.

5. Conclusion

In this paper, we proposed a effective and efficient method, called CDGNet, for human parsing. We exploited the pixel labeling of each category to produce the horizontal and vertical class distributions to reflect the distribution rule of each human part. With the guidance of the generated class distributions of each class in the horizontal and vertical directions, the network had more chance of learning the distribution consistent with the structural prior of the human body. It significantly benefited the labeling of each pixel in images with only one person and multiple persons as well. Extensive quantitative and qualitative comparisons indicated that the proposed CDGNet perform favorably against the state-of-the-art human parsing approaches.
Limitation: We assume that the number of classes is mostly about twenty (e.g., LIP, ATR, CIHP) and CDGNet can make the best accuracy. However, we note that CDGNet may not perform the best performance when using the small size of classes (e.g., one or two classes) in human parsing.
Potential Negative Societal Impact: CDGNet is a generic technology with many potential applications for human images. We are unaware of all potential works but can imagine that each application has its own merits and societal impacts depending on the intentions of the users.

References

[1] V. Badrinarayanan, A. Kendall, and R. Cipolla. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE TPAMI*, 39(12):2481–2495, Dec. 2017. 2

[2] L.-C. Chen, M. D. Collins, Y. Zhu, G. Papandreou, B. Zoph, F. Schroff, H. Adam, and J. Shlens. Searching for efficient multi-scale architectures for dense image prediction. *NeurIPS*, Dec. 2018. 2

[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. *IEEE TPAMI*, 40(4):834–848, 2017. 2, 6, 7

[4] L.-C. Chen, G. Papandreou, F. Schroff, and H. Adam. Rethinking atrous convolution for semantic image segmentation. *arXiv preprint arXiv:1706.05587*, 2017. 1, 2

[5] L.-C. Chen, Y. Yang, J. Wang, W. Xu, and A. L. Yuille. Attention to scale: Scale-aware semantic image segmentation. In *CVPR*, pages 3640–3649. IEEE, Jun. 2019. 2, 6, 7

[6] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *ECCV*, pages 833–851, Sep. 2018. 2

[7] S. Choi, J. T.Kim, and J. Choo. Cars can't fly up in the sky: Improving urban-scene segmentation via height-driven attention networks. In *CVPR*, pages 9373–9383, Jun. 2020. 3

[8] J. Fu, J. Liu, H. Tian, Y. Li, Y. Bao, Z. Fang, and H. Lu. Dual attention network for scene segmentation. In *CVPR*, pages 3146–3154, Jun. 2019. 1, 2, 5, 7, 8

[9] J. Fu, J. Liu, Y. Wang, Y. Li, Y. Bao, J. Tang, and H. Lu. Adaptive context network for scene parsing. In *ICCV*, pages 6748–6757, Oct. 2019. 1, 2

[10] K. Gong, Y. Gao, X. Liang, X. Shen, M. Wang, and L. Lin. Graphonomy:universal human parsing via graph transfer learning. In *CVPR*, page 7450–7459, Jun. 2019. 5, 6, 7

[11] K. Gong, X. Liang, Y. Li, Y. Chen, M. Yang, and L. Lin. Instance-level human parsing via part grouping network. In *ECCV*, pages 805–822, Sep. 2018. 5, 7

[12] K. Gong, X. Liang, D. Zhang, X. Shen, and L. Lin. Look into person: Self-supervised structure-sensitive learning and a new benchmark for human parsing. In *CVPR*, pages 932–940. IEEE, Jul. 2017. 5

[13] K. He, G. Gkioxari, P. Dollar, and R. Girshick. Mask r-cnn. In *CVPR*, pages 770–778. IEEE, Jun. 2016. 6

[14] Q. Hou, L. Zhang, M.-M. Cheng, and J. Feng. Strip pooling: Rethinking spatial pooling for scene parsing. In *CVPR*, pages 4003–4012, Jun. 2020. 1, 2

[15] J. Hu, L. Shen, and G. Sun. Squeeze-and-excitation networks. In *CVPR*, pages 7132–7141, Jun. 2018. 2

[16] L. Huang, Y. Yuan, J. Guo, C. Zhang, X. Chen, and J. Wang. Interlaced sparse self-attention for semantic segmentation. *arxiv preprint arXiv:1907.12273*, 2019. 2

[17] Z. Huang, C. Wang, X. Wang, W. Liu, and J. Wang. Semantic image segmentation by scale-adaptive networks. *IEEE TIP*, 29:2066–2077, 2020. 2

[18] Z. Huang, X. Wang, L. Huang, C. Huang, Y. Wei, and W. Liu. Ccnet: Criss-cross attention for semantic segmentation. In *ICCV*, pages 603–612, Oct. 2019. 1, 2

[19] R. Ji, D. Du, L. Zhang, L. Wen, Y. Wu, C. Zhao, F. Huang, and S. Lyu. Learning semantic neural tree for human parsing. In *ECCV*, pages 205–221, Aug. 2020. 2, 5, 6, 7

[20] Z. Jin, T. Gong, D. Yu, Q. Chu, J. Wang, C. Wang, and J. Shao. Mining contextual information beyond image for semantic segmentation. In *ICCV*, Oct. 2021. 2, 7

[21] Z. Jin, B. Liu, Q. Chu, and N. Yu. Isnet: Integrate image-level and semantic-level context for semantic segmentation. In *ICCV*, Oct. 2021. 1, 5, 7, 8

[22] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *NIPS*, pages 1097–1105, Dec. 2012. 6

[23] P. Li, Y. Xu, Y. Wei, and Y. Yang. Self-correction for human parsing. *IEEE TPAMI*, 2020. 6, 7, 8

[24] X. Liang, G. Ke, X. Shen, and L. Liang. Look into person:joint body parsing and pose estimation network and a new benchmark. *IEEE TPAMI*, 41(4), 2019. 6

[25] 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 *ICCV*, pages 1386–1394. IEEE, Dec. 2015. 5, 7

[26] J. Long, E. Shelhamer, and T. Darrell. Fully convolutional networks for semantic segmentation. In *CVPR*, pages 3431–3440, Jun. 2015. 1

[27] X. Luo, Z. Su, J. Guo, G. Zhang, and X. He. Trusted guidance pyramid network for human parsing. In *ACM MM*, pages 654–662, Oct. 2018. 7

[28] Y. Luo, Z. Zheng, Z. Liang, G. Tao, J. Yu, and Y. Yi. Macro-micro adversarial network for human parsing. In *ECCV*, pages 384–401, Sep. 2018. 6

[29] O. Ronneberger, P. Fischer, and T. Brox. U-net: Convolutional networks for biomedical image segmentation. In *MICCAI*, pages 234–241, Oct. 2015. 2

[30] T. Ruan, T. Liu, Z. Huang, Y. Wei, S. Wei, Y. Zhao, and T. Huang. Devil in the details: Towards accurate single and multiple human parsing. In *AAAI*, pages 4814–4821, 2019. 1, 2, 4, 5, 6, 7

[31] W. Wang, Z. Zhang, S. Qi, J. Shen, Y. Pang, and L. Shao. Learning compositional neural information fusion for human parsing. In *ICCV*, pages 5703–5713. IEEE, Oct. 2019. 2

[32] W. Wang, H. Zhu, J. Dai, Y. Pang, J. Shen, and L. Shao. Hierarchical human parsing with typed part-relation reasoning. In *CVPR*, pages 8929–8939, Jun. 2020. 2, 3, 6, 7, 8

[33] X. Wang, R. Girshick, A. Gupta, and K. He. Non-local neural networks. In *CVPR*, pages 7794–7803, Jun. 2018. 2, 5

[34] Y. Wang, J. Zhang, M. Kan, S. Shan, and X. Chen. Self-supervised equivariant attention mechanism for weakly supervised semantic segmentation. In *CVPR*, pages 12275–12284, Jun. 2020. 2, 5, 6

[35] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon. Cbam: Convolutional block attention module. In *ECCV*, pages 3–19, Sept. 2018. 3

[36] C. Yu, J. Wang, C. Gao, G. Yu, C. Shen, and N. Sang. Context prior for scene segmentation. In *CVPR*, pages 12416–12425, Jun. 2020. 1, 5

[37] Y. Yuan, X. Chen, and J. Wang. Object-contextual representations for semantic segmentation. In *ECCV*, pages 173–190, Aug. 2020. 1, 2

[38] Y. Yuan, L. Huang, J. Guo, C. Zhang, X. Chen, and J. Wang. Ocnet:object context network for scene parsing. *arXiv preprint arXiv:1809.00916*, 2018. 1, 2

[39] F. Zhang, Y. Chen, Z. Li, Z. Hong, J. Liu, F. Ma, J. Han, and E. Ding. Acfnet: Attentional class feature network for semantic segmentation. In *ICCV*, pages 6797–6806, Oct. 2019. 2

[40] L. Zhang, X. Li, A. Arnab, K. Yang, Y. Tong, and P. H. Torr. Dual graph convolutional network for semantic segmentation. In *BMVC*, pages 371–389, Sep. 2019. 2

[41] X. zhang, Y. Chen, B. Zhu, J. Wang, , and M. Tang. Part-aware context network for human parsing. In *CVPR*, page 8971–8980, Jun. 2020. 6, 7

[42] X. Zhang, Y. Chen, B. Zhu, J. Wang, and M. Tang. Blended grammar network for human parsing. In *ECCV*, pages 189–205, Aug. 2020. 2, 7

[43] Z. Zhang, C. Su, L. Zheng, and X. xie. Correlating edge, pose with parsing. In *CVPR*, pages 8900–8909, Jun. 2020. 2, 4, 5, 6, 7

[44] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia. Pyramid scene parsing network. In *CVPR*, pages 7450–7459, Jun. 2019. 1, 2

[45] J. Zhao, J. Li, X. Nie, F. Zhao, Y. Chen, Z. Wang, J. Feng, and S. Yan. Self-supervised neural aggregation networks for human parsing. In *CVPRW*, pages 4321–4329, Jul. 2017. 6

[46] Z. Zhu, M. Xu, S. Bai, T. Huang, and X. Bai. Asymmetric non-local neural networks for semantic segmentations. In *ICCV*, pages 593–602, Oct. 2019. 2, 5