3-2022

Mask-guided deformation adaptive network for human parsing

Aihua MAO
South China University of Technology

Yuan LIANG
South China University of Technology

Jianbo JIAO
University of Oxford

Yongtuo LIU
South China University of Technology

Shengfeng HE
Singapore Management University, shengfenghe@smu.edu.sg

Citation

MAO, Aihua; LIANG, Yuan; JIAO, Jianbo; LIU, Yongtuo; and HE, Shengfeng. Mask-guided deformation adaptive network for human parsing. (2022). *ACM Transactions on Multimedia Computing, Communications and Applications*. 18, (1), 1-20.
Available at: https://ink.library.smu.edu.sg/sis_research/7873

Mask-Guided Deformation Adaptive Network for Human Parsing

AIHUA MAO and YUAN LIANG, South China University of Technology, China
JIANBO JIAO, University of Oxford, United Kingdom
YONGTUO LIU and SHENGFENG HE, South China University of Technology, China

Due to the challenges of densely compacted body parts, nonrigid clothing items, and severe overlap in crowd scenes, human parsing needs to focus more on multilevel feature representations compared to general scene parsing tasks. Based on this observation, we propose to introduce the auxiliary task of human mask and edge detection to facilitate human parsing. Different from human parsing, which exploits the discriminative features of each category, human mask and edge detection emphasizes the boundaries of semantic parsing regions and the difference between foreground humans and background clutter, which benefits the parsing predictions of crowd scenes and small human parts. Specifically, we extract human mask and edge labels from the human parsing annotations and train a shared encoder with three independent decoders for the three mutually beneficial tasks. Furthermore, the decoder feature maps of the human mask prediction branch are further exploited as attention maps, indicating human regions to facilitate the decoding process of human parsing and human edge detection. In addition to these auxiliary tasks, we further alleviate the problem of deformed clothing items under various human poses by tracking the deformation patterns with the deformable convolution. Extensive experiments show that the proposed method can achieve superior performance against state-of-the-art methods on both single and multiple human parsing datasets. Codes and trained models are available https://github.com/ViktorLiang/MGDAN.

CCS Concepts: • **Computing methodologies → Image segmentation**; **Appearance and texture representations;**

Additional Key Words and Phrases: Human parsing, multi-task learning, deformable convolution

A. Mao and Y. Liang contributed equally to this research.

This project is supported by the National Natural Science Foundation of China under Grant No.:~61972162; Guangdong International Science and Technology Cooperation Project (No. 2021A0505030009); Guangdong Natural Science Foundation (No. 2019A1515010833, 2021A1515012625); Guangzhou Basic and Applied Research Project (No. 202102021074); the Fundamental Research Funds for the Central Universities (No. 2020ZYGXZR089); the Social Science Research Base of Guangdong Province-Research Center of Network Civilization in New Era of SCUT; and the CCF-Tencent Open Research fund under Grant No.: CCF-Tencent RAGR20190112.

Authors' addresses: A. Mao, Y. Liang, Y. Liu, and S. He (corresponding author), School of Computer Science and Engineering, South China University of Technology, Guangzhou, 510006, China; emails: ahmao@scut.edu.cn, yuanliang07@gmail.com, csmanlyt@mail.scut.edu.cn, hesfe@scut.edu.cn; J. Jiao, Department of Engineering Science, University of Oxford, Parks Road, Oxford, OX1 3PJ, United Kingdom; email:jianbo@robots.ox.ac.uk.

1551-6857/2022/03-ART11 $15.00
https://doi.org/10.1145/3467889

ACM Reference format:
Aihua Mao, Yuan Liang, Jianbo Jiao, Yongtuo Liu, and Shengfeng He. 2022. Mask-Guided Deformation Adaptive Network for Human Parsing. *ACM Trans. Multimedia Comput. Commun. Appl.* 18, 1, Article 11 (March 2022), 20 pages.
https://doi.org/10.1145/3467889

1 INTRODUCTION

Human parsing is a segmentation task for fine-grained human body parts and clothing items, which aims at assigning each human-related pixel a semantic label. Human parsing can benefit several human-centric tasks such as person re-identification [17, 26, 34, 44], human behavior recognition [21, 37], and clothing fashion retrieval [23, 48].

State-of-the-art human parsing approaches mainly focus on exploiting rich prior knowledge to improve parsing performance. Among them, edge detection is introduced by [2, 10, 36, 39] as a boundary prior to assist human parsing. Despite the promising performance, edge and parsing results of small-scale objects tend to be coarsely predicted or mispredicted as background without the high-level guidance of foreground and background constraints. As a result, in crowded scenes where the close-to-camera humans have a rather larger scale than those off-camera, the small-scale humans are easily ignored by current models (see Figure 1).

In addition to the lack of high-level constraints, the problem of background/foreground misprediction is also due to the data imbalance of human parsing annotations. Statistically, more than 50% of pixels are labeled as background in the human parsing dataset LIP [11], while the rest of the samples are exploited to recognize 19 categories of human body parts. This heavily biases the network training toward the easy background samples, resulting in inaccurate predictions for foreground human body parts and clothing items.

A recent line of work [9, 12, 16, 40–42] proposes to manually group the parsing labels into several pyramid levels and elucidate the hierarchical adjacent relationships between human body parts via Graph Neural Networks or part-relation reasoning. Another line of work [18, 36] leverages binary edge detection to embed the local part relations in edge feature maps. The edge feature maps provide a simple but efficient part-relation reasoning since only binary edge prediction is generated.

To address the aforementioned shortcomings and combine their respective merits, we propose to integrate human edge and mask detection simultaneously to fully exploit the mutually beneficial low-level and high-level constraints for accurate parsing estimations. In a similar vein to the hierarchical structured human parsing methods [9, 12, 16, 40–42], our proposed mask prediction branch first generates full-body features and then decomposes them into human parsing features. Like SCHP [18], our method also incorporates low-level edge information. Unlike them, the edge features are served as supplemented details, which are aggregated with both mask features and parsing features to generate final parsing prediction. To the best of our knowledge, it is the first attempt to explore the fusion strategy of multilevel constraints achieved by simultaneous human edge and mask detection for human parsing.

Another challenge for human parsing is to distinguish different kinds of body parts and clothing items under various human poses. However, such deformation problem in human parsing is still not well explored in the literature. Recent methods [6, 31] introduce auxiliary human pose constraints to alleviate this problem, but deformed clothing may not always be consistent with the rigid human body parts and presents more complex variations due to the nonrigid nature. To this end, inspired by the success of deformable convolutions, we propose to explore them to adaptively learn intrinsic representations of body and clothing deformations.

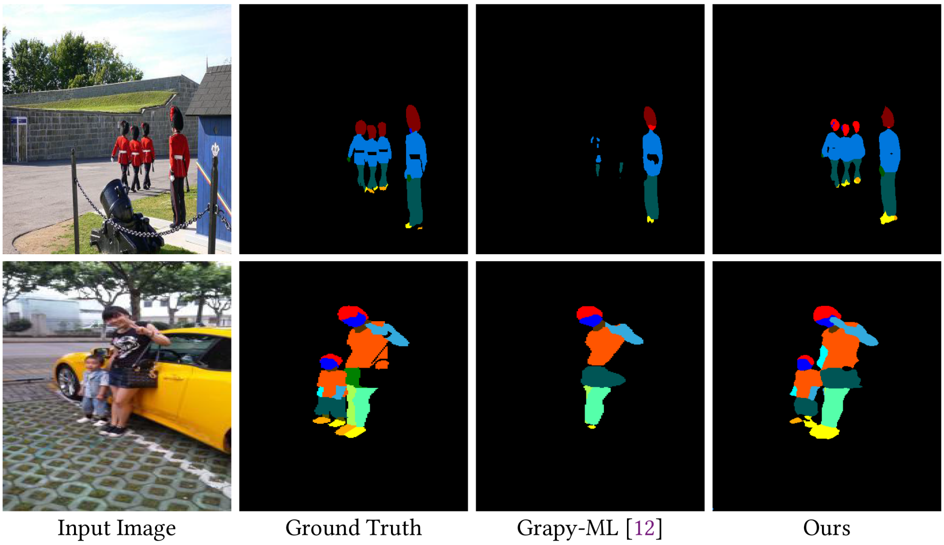


Fig. 1. Two challenging examples of human parsing on the CIHP [10] datasets. This illustrates crowd scenes, where small-scale individuals in the off-camera regions are ignored by Grapy-ML [12]. Differently, our method can alleviate this issue by introducing auxiliary tasks of human mask and edge detection facilitated with deformable convolutions.

To summarize, we propose a mask-guided deformation adaptive network for human parsing. Specifically, in addition to the primary human parsing branch, we introduce the auxiliary task of simultaneous human mask and edge detection. The human mask prediction branch ignores specific categories of human body parts and concentrates more on the high-level foreground and background separations. The aforementioned data imbalance problem can be elegantly alleviated as the foreground or background class contains roughly the same amount of training samples (pixels). The human edge detection branch is dedicated to the low-level boundary calibrations between adjacent parsing items. It can provide detailed information regarding small-scale human body parts, of which human parsing tends to have relatively low prediction confidence. The mutually beneficial human understanding tasks share the same feature extractor and utilize three different decoders to output human parsing, human mask, and edge detection, respectively. Additionally, the decoder feature maps of the two auxiliary branches are further exploited to facilitate the decoding process of human parsing. Specifically, the decoded mask feature maps are multiplied with those of the human parsing branch to serve as attention maps, indicating human regions for better discriminating foreground humans and background clutter. Meanwhile, the decoded edge feature maps are shared with the human parsing branch via concatenation to emphasize accurate boundaries of multiscale human body parts. More importantly, the backbone network is not only governed by human parsing but also deeply supervised by human mask and edge detection in a multitask fashion. This design releases the training burden of foreground/background separation and boundary calibration from human parsing and allows the learning of human parsing to focus more on human body part classification. Furthermore, as the vanilla convolutional layer is performed via a grid-structured kernel, which limits the capabilities of capturing shape variations caused by human pose changes, we cope with this problem by introducing **deformable convolution (DCN)** [56]. Specifically, we utilize deformable convolutions in the last three blocks of the backbone network in order to deal with the deformation in a multiscale fashion. Extensive experiments are conducted to explore the fusion strategies of these mutually beneficial tasks and demonstrate superior performance against state-of-the-art methods. In summary, the main contributions of our work are threefold:

- We propose to integrate human edge and mask detection simultaneously to fully exploit the mutually beneficial low-level and high-level constraints for human parsing. This reduces the learning ambiguity of human parsing by disentangling this task into foreground-background separation, boundary calibration, and human body part classification, respectively, and therefore substantially reduces the parsing prediction errors.
- We propose to introduce deformable convolutions to adaptively learn intrinsic representations of deformed body parts and clothing items induced by various human pose changes.
- Extensive experiments show that the proposed method can achieve superior performance with respect to three widely used benchmarks of both single-human and multihuman parsing.

2 RELATED WORK

Here we discuss related literature in three aspects: human parsing, scene parsing, and object deformation modeling.

2.1 Human Parsing

Great interest has long been attracted to human parsing in the literature. Traditional human parsing methods [23, 27, 29, 48] utilize hand-designed over-segmentation (i.e., HOG, superpixels) and hand-crafted structures (i.e., And-Or graph) to build models [43, 45], and unsupervised superpixels and CRF are commonly used to refine predicted labels. Generally, the pipelines are composed of several hand-designed models, which readily leads to bottlenecks among them. Early **convolutional neural network (CNN)**-based approaches [24, 28] utilize feature extraction and region relation learning combined in an end-to-end manner. Although much improvement has been achieved, the limited CNN layers cannot learn abundant discriminative feature representations. With the proposal of deep residual networks [13], deeper parsing features boost the performance of various human parsing approaches [22, 30, 53, 55]. In spite of the improved feature extraction, some problems like the prediction of small-scale human body parts, deformed clothes, and crowd scenes still remain challenging for human parsing. Recently, several methods [30, 53, 55] are proposed to tackle the scale variation problem. Zhao et al. [53] employ three versions of each input image with scaling factors of 0.5, 0.75, and 1, and each version is processed by a fully CNN with shared weights to learn the scale-invariant feature representations.Some works [10, 36] utilize pyramid pooling [52] in the top layers to abstract multilevel contextual representations for human parsing. In addition, they also introduce an auxiliary human edge detection branch. Although parsing boundaries can be calibrated to some extent in such methods, edge and parsing results of small-scale objects tend to be coarsely mispredicted as background without the high-level guidance of foreground and background constraints.

Inspired by the human visual system, Zhu et al. [55] design a hierarchical structure, where large-scale human body parts are predicted at lower layers and serve as prior information for small-scale human body parts that are predicted at deeper layers. To further explore the inherent hierarchical structure of a human body, BGNet [50] leverages a grammar rule to first predict conspicuous parts (e.g., torso, head), which progressively amends the prediction of inconspicuous parts (e.g., low-arm, low-leg). Some of the other latest works explore the inherent structure of human bodies and propose to manually group the parsing labels into several pyramid levels (e.g., full-body, upper-lower body, detailed parsing body) [9, 12, 40–42] or cascade tree structure [16]. To constrain the structured human parsing results, they typically formulate the adjacent relationships between human body parts via graph neural networks [9] or part-relation reasoning [12, 16, 40–42]. In addition to the network architecture design and intrinsic consistency constraints, several works [18, 19] are specialized to deal with the problem of insufficient high-quality labels. To remedy the problem of noisy labels, SCHP [18] proposes to train the model by online aggregation and refine the noisy labels synchronously. To alleviate the problem of insufficient training samples, Li et al. [19] design a self-learning strategy for efficient supervision. Our model design follows the same spirit as hierarchical structure models [9, 12, 16, 40–42], with the complementary refinement by aggregating three predictions with learned global convolutions.

2.2 Scene Parsing

As a related task to human parsing, scene parsing also suffers from the issues caused by object scale variations. Aiming at obtaining more global and local scene category clues, Zhao et al. [52] first propose pyramid pooling with several pooling layers executed independently to obtain multiscale features. Fu et al. [8] employ self-attention modules with spatialwise nonlocal module and channelwise SEResNet module [14] to integrate contextual features. To learn the distinct features of foreground and background, some methods [20, 47] are proposed to utilize two branches for foreground and background predictions individually. Li et al. [20] propose to leverage foreground features as an attention map to guide the background branch by conducting RoI Upsampling on the learned features of bounding boxes. In [47], the foreground and background features are fused in the panoptic head to obtain the final dense prediction. Our method shares the same spirit with these methods targeting reduction of the learning ambiguities of the major task. Differently, we tailor the network for human parsing by introducing two auxiliary tasks to aid the primary parsing task.

2.3 Deformation Modeling

Object deformation caused by viewpoint changes or nonrigid transformation has always been a challenge in visual recognition. Felzenszwalb et al. [7] propose to consider the mixture of root filter and part filters, where the former responds for coarse detection of an entire object, and the latter is utilized to handle detailed deformations of each part. The filters are learned by a latent SVM and the detection score is a combination of root and part responses. Jaderberg et al. [15] first utilize the CNN to tackle object deformations by a spatial transformer that warps the input features by a learnable affine transformation matrix. This matrix is applied isotropically where features at different channels share the same affine transformation, which limits its ability to channel-sensitive tasks, such as semantic segmentation, where more dense or semi-dense predictions are needed. Instead of spatially adjusting the features by a learnable transformation matrix, Dai et al. [5] handle deformation by learning sample offsets for regular convolution kernels. These learnable offsets enlarge the sample range of each convolution kernel in a local and dense manner, which makes it more appropriate for complex tasks. In [56], an improved deformable version with the modulation mechanism is proposed to adaptively adjust offsets and modulate the input feature amplitudes. Differently, we adopt deformable convolution to alleviate the deformed regions of human items, especially clothing. However, the learned offsets tend to span background regions between pairwise body parts such as arms and legs; thus, we further introduce a human mask detection branch serving as a background filter for rectification.

3 APPROACH

In this section, we present our approach in detail. As aforementioned, our method mainly addresses the foreground and background imbalance problem along with deformation issues in human

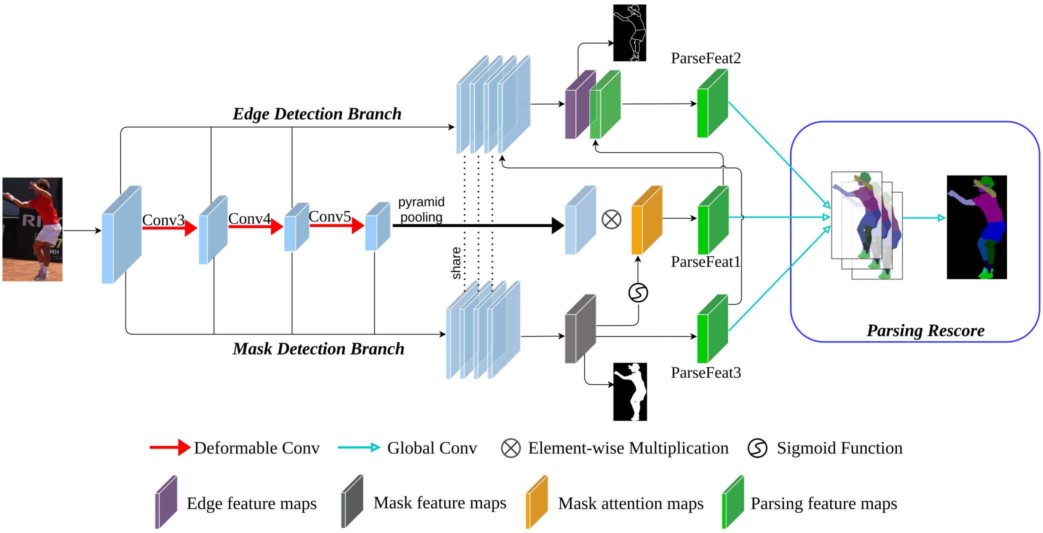


Fig. 2. Illustration of our proposed method. In addition to the primary human parsing architecture, two auxiliary branches are introduced to explicitly capture the multilevel prior context of human mask and edge detection. Specifically, the human mask prediction branch guides human parsing by attentive multiplication in the decoding process, while the edge branch provides detailed boundary priors to facilitate fine-grained parsing predictions. Finally, feature maps from the three branches (e.g., ParseFeat1, ParseFeat2, and ParseFeat3) are fused by concatenation and convolution to refine the final parsing result.

parsing. As shown in Figure 2, the pipeline of the proposed model mainly consists of four parts: human parsing branch, human mask prediction branch, human edge detection branch, and the parsing rescoring module. Specifically, the backbone network facilitated with deformable convolution layers (e.g., red arrows in Figure 2) aims to learn deformation adaptive feature representations that are fed to the three downstream mutually complementary tasks. The mask prediction branch leverages multiscale features from the backbone network to provide human mask attentions for human parsing. The attention maps are also explored as foreground prior knowledge for the edge detection branch, which facilitates parsing predictions of small-scale object categories. Finally, parsing predictions from the three branches are processed by a rescoring module to refine the final human parsing result while taking their complementary merits into consideration.

3.1 Backbone Network

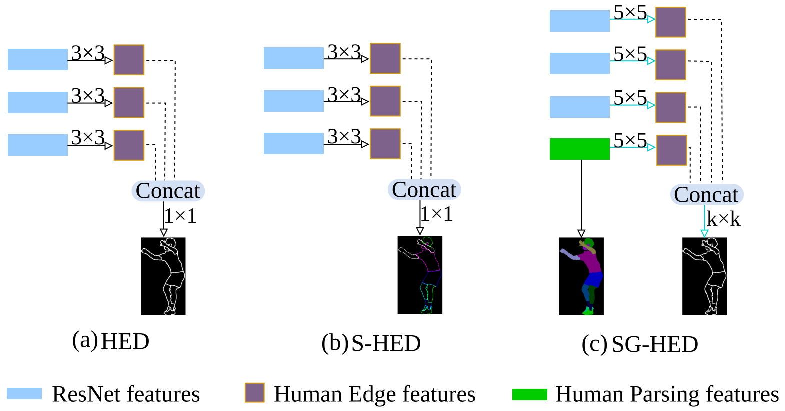
We modify ResNet-101 [13] as our backbone network. To alleviate the limitations of grid-structured kernels of vanilla convolutional layers, we employ DCN [56] in the deeper layers of the backbone network.

Denoting the sampling grid of a convolutional layer as $\mathcal{R}$ for the input feature map x, the output feature map y at each location p_0 can be formulated as

$$y(p_0) = \sum_{p_n \in \mathcal{R}} w(p_n) \cdot x(p_0 + p_n). \tag{1}$$

With the learnable offset $\triangle p_n$ for each sampled location, the result of deformable convolution can be formulated as

$$y(p_0) = \sum_{p_n \in \mathcal{R}} w(p_n) \cdot x(p_0 + p_n + \triangle p_n). \tag{2}$$



(a) HED (b) S-HED (c) SG-HED

Fig. 3. Three types of designs for human edge detection. For the space limitation, "HED," "S-HED," and "SG-HED" indicate human edge detection, semantic human edge detection, and semantic-guided human edge detection, respectively. (a) is the vanilla setting where human edge labels are extracted from the human parsing annotations without regard to semantic categories of each human body part. (b) shows the semantic human edge detection [49] that allocates human parsing categories to extracted human edges. (c) is the proposed method modified from (a) and (b) to explicitly guide human edge detection with semantic constraints from human parsing.

The learnable offset $\triangle p_n$ provides an adaptive selection strategy for tracking the deformation patterns of clothing items under various human poses. In our model, the convolutional layers from the third residual block to the last one (e.g., Conv3, Conv4, and Conv5 in Figure 2) are replaced by DCN. It is worth noting that low-level features contain more accurate object locations that benefit parsing detailed object boundaries; thus, multilevel features from the backbone network are integrated to generate the initial discriminative feature maps for human parsing (e.g., ParseFeat1 in Figure 2).

3.2 Human Edge Detection Branch

Human edges serve as boundary prior to distinguish highly compact clothing items and body parts in human parsing. Different from the general edge detection problem [46], which aims at finding all salient edges, edge detection for human parsing (see Figure 3(a)) is mainly dedicated to discriminating human edges between various clothing items and body parts, which relies more on top-down semantic guidance. To explore the optimal settings for the human edge detection branch, we attempt to utilize the semantic human edge detection [49] as shown in Figure 3(b), where each edge has a semantic label. However, we find that the semantic human edge detection struggles to converge to what we want it to learn, and hardly provides general boundary constraints for human parsing. To this end, we exploit another variant, namely, semantic-guided human edge detection (see Figure 3(c)), which utilizes human parsing as semantic guidance for human edge detection. Specifically, as shown in Figure 2, multiscale feature maps extracted from the backbone network are processed by a 1×1 convolutional layer to generate three groups of multilevel feature maps with the same resolution and each having 256 channels. ParseFeat3, which originates from multilevel combined features, is transferred to the human edge detection branch for high-level foreground and background constraints, and the transferred features together with the multilevel combined features are both supervised by the human parsing and human edge detection branches, which effectively bridges the multilevel mutually complementary human understanding tasks.

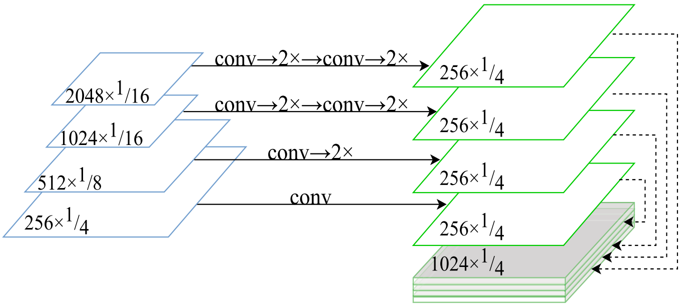


Fig. 4. Illustration of the fusion process for multilevel combined feature maps. Each group of feature maps from the backbone network is processed via convolutional layers and optional bilinear upsampling operations to 1/4 size of the input image.

Denoting the four groups of feature maps in the edge detection branch as P_{edg1}, P_{edg2}, P_{edg3}, and P_{edg4}, the final edge prediction P_{edg} is formulated as

$$P_{edg} = f_k(Concat(P_{edg1}, P_{edg2}, P_{edg3}, P_{edg4})), \tag{3}$$

where *Concat* means concatenation along the channel dimension, and f_k represents a convolutional layer with kernel size of k. In addition, to further exploit the low-level boundary constraints, we propose to divide the feature maps (i.e., P_{edg1}, P_{edg2}, P_{edg3}, and P_{edg4}) into grids (nonoverlapping patches) and concatenate them along the batch dimension before they are fed into the following convolutional layers. The rationale behind this design is that boundary calibrations can be realized from a local perspective, and the grid-dividing method can confine the receptive field and be regarded as a self-supervised learning strategy for human edge detection. In experiments, we empirically set the grid as 2×2. The quantitative results of our method with grid division in the edge detection branch are marked as "Ours*" in Tables 7, 8, and 9, where "Ours" denotes the variant without this technique.

3.3 Human Mask Prediction Branch

Human mask prediction also benefits from multilevel feature representations [33, 54]. Therefore, derived from the backbone network, we build a new branch that extracts multilevel features for human mask prediction. Similar to the edge detection branch, the outputs from Conv2, Conv3, Conv4, and Conv5 are processed by convolutional layers followed by necessary upsampling operations (or not) to generate four groups of feature maps with the same resolution. See Figure 4 for a detailed configuration. Aiming at providing soft attention to highlight foreground regions, the integrated features are processed by the sigmoid function and then multiplied with the feature maps in the human parsing branch.

The full-body supervised mask branch tends to smooth the boundary details between different parsing categories, and thus the mask-guided full-body attentions have the potential to undermine the activation in parsing features. To further amend this problem, we integrate mask feature maps for human edge detection and human parsing prediction. Besides, the input features from the backbone network for mask and edge detection branches are also shared. With our full-body attention, the high activations of edge features are restricted in foreground human regions. The mask features, on the other hand, could preferably keep the desired activations in the boundary regions, which might be smoothed by the mask supervision. The effectiveness of combined features as attention maps for the human parsing branch can be better guaranteed when the combined features are supervised by both edge detection and mask prediction branches. The effectiveness of this fusion strategy is demonstrated in Table 2 and visualized in Figure 5.

Formally, we denote the concatenated four groups of feature maps in the mask branch as F_{cat}, and then the multilevel combined features F_{mask} generated before the sigmoid activation can be formulated as

$$F_{mask} = \sigma(f_3(F_{cat})), \tag{4}$$

where f_3 is a convolutional layer with kernel size of 3, and σ represents the ReLU activation function. For simplicity, we denote $ParseFeat1$ as F_{ps1}, with the feature maps generated by Conv3 and multiscale feature pyramid pooling denoted as F_{conv3} and F_{psp}, respectively. F_{ps1} can then be formulated as

$$F_{ps1} = f_1(f_1(F_{conv3}), F_{psp}) * sigmoid(F_{mask}), \tag{5}$$

where f_1 denotes a convolutional layer with kernel size of 1, and *sigmoid* means the sigmoid activation function.

3.4 Human Parsing Rescoring Module

For simplicity, we denote multilevel combined features from the three branches as P_{ps1}, P_{ps2}, and P_{ps3} (corresponding to ParseFeat1, ParseFeat2, and ParseFeat3 in Figure 2). These three groups of feature maps are obtained from three mutually complementary tasks and contain rich multiscale and multilevel feature representations. Specifically, P_{ps1} is the combination of multiscale feature maps derived from the backbone network, which are elementwisely multiplied by P_{ps3} to attentively discriminate human body parts from background clutter, while P_{ps2} is supervised by both the human parsing branch and the edge detection branch, which can facilitate the discoveries of boundaries and estimation of more accurate parsing results.

Appropriately incorporating the bright sides of the three branches can further benefit estimations of challenging small-scale categories and constrain the local consistency for large-scale items. To this end, neighboring pixels should also be taken into consideration for inference. For instance, if we want to judge whether one pixel is background or not, the prediction results of its neighbors should also be taken into consideration, and the wider the better. Thus, we propose a parsing rescoring module with a large receptive field for the combination of P_{ps1}, P_{ps2}, and P_{ps3}. By integrating multibranch combined features for parsing, the results are further refined by the rescoring module. Without loss of generality, we simply implement it as a convolutional layer.

Denoting the result from the parsing rescoring module as $P_{ps\_rescore}$ and a convolutional layer with kernel size of k as f_k, the overall parsing prediction $P_{ps\_rescore}$ is formulated as

$$P_{ps\_rescore} = f_k(Concat(P_{ps1}, P_{ps2}, P_{ps3})). \tag{6}$$

Note that P_{ps2} is the final parsing prediction, while other predictions serve as supervisors to enforce guidance for better-learned features. As shown in Table 5, slightly better performance improvement has been achieved with the rescore module, especially on the multihuman dataset CIHP.

3.5 Training Details

The labels for the human mask prediction branch are generated by assigning all nonbackground pixels as 1 and background pixels to 0, while human edge labels are extracted in the same way as [36]. Cross-entropy is utilized as the loss function for all pixelwise predictions:

$$L = -\sum_{c=1}^{N} w_c y_c log(p_c), \tag{7}$$

where N denotes the number of categories, which is 20 for the LIP and CIHP datasets, and 2 for the edge and mask branches. w_c is the balance weight for category c. In experiments, we set all w_c to 1 for all categories except the edge detection branch. For edge detection, we set the nonedge ratio as the balance weight for the edge category, while the edge ratio is set as the balance weight for the nonedge category. Since the ratio of background and foreground is naturally balanced across the datasets, we also set balance weights to 1 for simplicity in the human mask prediction branch.

Denoting the loss functions for human parsing results of the three branches and the downstream rescoring module as L_{ps1}, L_{ps2}, L_{ps3}, and $L_{ps\_rescore}$, respectively, the overall loss function L_{ps} of the human parsing branch can be formulated as

$$L_{ps} = L_{ps1} + L_{ps2} + L_{ps3} + L_{ps\_rescore}. \tag{8}$$

Each branch has a shared top-level feature from the backbone. The loss values between any parsing predictions are not differing much. Thus, each loss weight is set to 1.

By denoting loss functions for human mask prediction and edge detection as L_m, L_{edg}, the total loss function is formulated as

$$L_{total} = L_{ps} + L_{edg} + L_m. \tag{9}$$

The edge detection branch and the mask prediction branch are both supervised by human parsing prediction. Loss values for three tasks have no order-of-magnitude difference. Thus, we also set the loss weights for three tasks to 1.

Our network is optimized by SGD [38]. The initial learning rate is set to 0.001 and decayed by $lr \times (\frac{1 - iter}{total\_iter})^{0.9}$. The network is trained for 200 epochs with batch size of 21. The fixed resolution of input images is 384 × 384 for LIP, and 448 × 448 for CIHP and Pascal-Person-Part. Center cropping, scaling, flipping, and grayscaling with a probability of 0.5 are used for data augmentation.

4 EXPERIMENTS

In this section, we present extensive experiments to demonstrate the effectiveness of the proposed method.

4.1 Dataset and Evaluation Metric

We conduct comparison experiments on the following three datasets.

LIP: The LIP [11] dataset is the largest existing single human parsing dataset, with 50,462 images split into 30K/10K/10K for training, validation, and testing, respectively. The dataset contains 19 semantic labels for human body parts and clothing items. Human poses vary greatly in this dataset.

CIHP: The CIHP [10] dataset focuses on multihuman parsing and contains 38,280 images among 20 categories. Each image has semantic human parsing, instance-level human parsing, and human pose annotations. The dataset is split into 28,280/5,000/5,000 for training/validation/testing. Compared with other multihuman parsing datasets, CIHP exhibits higher resolutions and more crowd scenarios.

PASCAL-Person-Part: This dataset is constructed from the person annotations of the PASCAL-Part dataset [4]. The PASCAL-Part dataset provides pixel annotations for parts of animals, humans, vehicles, and so forth. The PASCAL-Person-Part dataset takes human annotations from it and groups them into six human body parts, e.g., head, torso, upper arms, lower arms, upper legs, and lower legs. This dataset contains 3,535 images, which are split into 1,717/1,818 for training/testing. Multiple humans and large-scale variations of human sizes could be found in this dataset.

For measuring the performance, the parsing accuracy is evaluated by mean IoU, pixel accuracy, and mean accuracy, in which mean IoU is the main metric to compare different human parsing

Table 1. Comparison Results of Different Configurations When Replacing the Convolutional Kernels in the Backbone Network by Deformable Convolution (DCN)

Method	pixel acc.	mean acc.	mIoU
CE2P[36]	87.37	63.2	53.1
CE2P+DCN001	86.55	64.01	51.68
CE2P+DCN011	87.46	**66.65**	54.19
CE2P+DCN110	87.32	66.13	53.72
CE2P+DCN111	**87.64**	66.37	**54.49**

Experiments are performed on the LIP dataset. DCN-xyz denotes the block of conv3(x), conv4(y) and conv5(z), where x, y, and z take the values of 1 or 0 to indicate replacement or lack thereof.

models. To be consistent with previous works, each dataset is tested using the model trained on the corresponding training set only.

4.2 Evaluation of Deformable Convolution

To explore the influence of DCN, the baseline model is kept unchanged except for the modifications on convolutional layers from Conv3 to Conv5 of the backbone network. We gradually replace them with DCNs. The comparison results of different modification strategies are shown in Table 1.

It can be seen that only replacing the Conv5 layer with DCN results in a slight accuracy decrease. Marginal improvements can be found by replacement in lower levels of feature extraction (e.g., Conv3 and Conv4). With all convolutional layers from Conv3 to Conv5 replaced by DCNs, the optimal performance is achieved, with mIoU improved by more than 1.3 against the baseline model. This indicates that multilevel features need to be taken into consideration for generating sampling offsets due to multiscale human body parts and clothing items.

In the following experiments, DCNs are applied to Conv3, Conv4, and Conv5 by default if there are no extra declarations.

4.3 Evaluation of Mask Prediction

In Table 2, we explore different fusion strategies between the human mask prediction branch and human parsing. The term "w/o Mask" indicates the baseline model without the human mask prediction branch. When we add this branch and leverage mask features to serve as attention maps (denoted as "w/ Mask Atten") or communicate with the human edge detection branch (denoted as "w/ Mask Edge") in the decoding process, the parsing performance becomes slightly improved. When we integrate each strategy with additional human parsing supervision (denoted as "w/ Mask Atten+Parse" or "w/ Mask Edge+Parse"), the performance is improved accordingly, with overall promotion of all three metrics. This indicates that the supervision signal from human parsing can refine the decoded feature maps of the human mask prediction branch for compatible fusions with the other branches. In addition, we train another variant to integrate these three strategies (denoted as "w/ Mask Atten+Edge+Parse"). As seen in Table 2, the performance reaches the optimum, which demonstrates the effectiveness of the introduced human mask prediction branch and multiple fusion strategies.

In addition to the quantitative comparison, we visualize the attention maps (the gray block in Figure 2) of "w/ Mask Atten" and "w/ Mask Atten+Edge+Parse" in Figure 5. We can see that the attention maps of "w/ Mask Atten" mainly show high response values outside the boundaries of human body parts, and direct attentive multiplication with the human parsing branch may

Table 2. Comparison Results of Different Fusion Strategies between the Human Mask Prediction Branch and Human Parsing on the LIP Dataset

Method	pixel acc.	mean acc.	mIoU
w/o Mask	87.64	66.13	54.49
w/ Mask Atten	87.84	66.67	54.93
w/ Mask Edge	87.83	66.24	54.85
w/ Mask Atten+Parse	87.99	67.43	55.38
w/ Mask Edge+Parse	87.43	67.52	55.12
w/ Mask Atten+Edge+Parse	88.07	67.19	55.58
w/ Mask Atten+Edge+Parse+Share	**88.24**	**67.81**	**56.32**

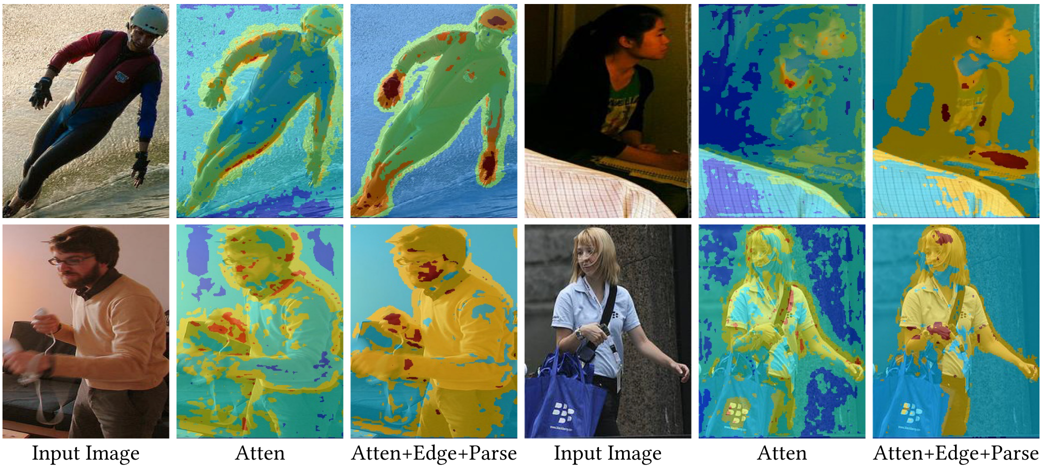


Fig. 5. Visualization of the attention maps extracted from the human mask prediction branch (marked by the gray block in Figure 2). Due to the space limitation, “Atten” and “Atten+Edge+Parse” represent “w/ Mask Atten” and “w/ Mask Atten+Edge+Parse,” respectively.

submerge prominent discriminative responses of interior boundaries. When equipped with the human edge and parsing supervision, the attention maps of “w/ Mask Atten+Edge+Parse” can render more regions of foreground humans, which maximizes the inherent strength of the human mask prediction branch for foreground/background separation and benefits the parsing predictions especially in crowded scenes. Furthermore, we share the weights of the backbone network among the three branches and show the result, which is denoted as “w/ Mask Atten+Edge+Parse+Share.” It is worth noting that the shared version can boost the performance of human parsing with fewer parameters than the unshared counterpart, which further verifies the effectiveness of the introduced mutually complementary tasks.

Apart from the averaged accuracy of all categories shown in Table 2, we also summarize the detailed performance of each parsing item using mIoU in Table 3. Compared with the baseline model “B,” the modified backbone network with deformable convolution (denoted as “B+D”) can achieve remarkable improvements for large-scale categories (e.g., *dresses*, *pants,* and *jumpsuits*) and easily deformed items (e.g., *gloves*, *scarves,* and *skirts*). Additionally, when introducing the human mask prediction branch (denoted as “B+D+M”), the small-scale categories (e.g., *socks*, *faces*, *legs,* and *shoes*), which perform inferiorly in “B+D,” can substantially benefit from the balanced foreground and background separations. Those infrequent categories, such as *hats*, *scarves,* and

Table 3. Detailed Comparison Results of Each Category on the LIP Dataset

Method	mIoU	bkg	hat	hair	glove	glasses	u-clothes	dress	coat	socks	pants	jmpsuits	scarf	skirt	face	l-arm	r-arm	l-leg	r-leg	l-shoe	r-shoe
B	53.1	87.67	65.29	72.54	39.09	32.73	69.46	32.52	56.28	49.67	74.11	27.23	14.19	22.51	75.5	65.14	66.59	60.1	58.59	46.63	46.12
B+D	54.49	88.05	67.27	72.14	41.14	32.38	70	35.42	56.82	48.63	75.58	34.2	22.31	31.50	74.61	65.53	67.76	57.97	57.22	45.33	45.96
B+D+M	55.38	88.45	67.11	72.74	45.35	31.49	70.84	35	57.93	50.67	76.15	33.95	22.27	28.37	75.25	66.84	69.09	60.07	59.7	47.72	48.58
B+D+M+R	55.84	88.45	67.67	72.93	45.08	32.27	70.98	36.27	58.1	51.23	76.29	35.24	23.94	28.55	75.42	66.94	69.36	61.07	60.23	47.84	48.85

B, *D*, *M,* and *R* denote the baseline model [36], deformable convolution, human mask prediction branch, and parsing rescoring module, respectively. Due to the space limitation, the best accuracy of each column is underlined.

Table 4. Detailed Comparison Results of Each Category with Respect to the Auxiliary Human Edge Detection Branch

Method	mIoU	bkg	hat	hair	glove	glasses	u-clothes	dress	coat	socks	pants	jmpsuits	scarf	skirt	face	l-arm	r-arm	l-leg	r-leg	l-shoe	r-shoe
Single human parsing on LIP																					
w/o Edge	55.0	88.17	66.14	71.90	43.19	30.84	70.89	36.08	58.04	49.03	75.61	34.43	24.30	28.86	74.45	65.29	68.22	60.34	60.11	46.52	48.05
w/ Edge	56.56	88.71	67.71	73.14	45.49	33.38	71.25	39.75	58.45	51.63	76.52	34.13	25.72	29.53	75.60	67.87	70.06	62.50	61.62	48.71	49.43
Multiple human parsing on CIHP																					
w/o Edge	61.8	94.21	71.32	80.99	33.48	57.02	69.42	58.67	66.07	36.10	74.82	73.03	35.85	43.15	87.88	70.58	71.55	60.04	60.45	47.32	45.93
w/ Edge	61.91	94.18	71.38	81.03	33.14	58.17	69.13	58.76	65.59	35.21	74.97	73.77	36.19	41.88	87.95	70.65	71.93	60.27	61.19	46.71	46.06

Variants of our method with and without the edge detection branch are denoted as “w/ Edge” and “w/o Edge,” respectively.

glasses, are still challenging for the introduced human mask prediction task. Therefore, adding the auxiliary human mask prediction branch may not boost the performance significantly for those categories. However, for the majority of parsing items, the human mask prediction branch can improve the performance through the introduced explicit semantic constraints. The last row entitled “B+D+M+R” represents an additional parsing rescoring module based on “B+D+M,” which will be discussed in Section 4.5.

4.4 Evaluation of Edge Detection

To evaluate the effectiveness of the human edge detection branch, we train two variants of the proposed method with and without the auxiliary branch, respectively denoted as “w Edge” and “w/o Edge” in Table 4. To be specific, “w/o Edge” represents the model where all the blocks and lines in Figure 2 specified for the edge detection branch are removed. As seen in Table 4, the introduced human edge detection branch can consistently improve the average accuracy (mIoU) for both single and multiple human parsing, which demonstrates the efficacy of low-level boundary constraints achieved by the auxiliary explicit human edge supervision. Compared with multiple human parsing on CIHP, single human parsing on LIP exhibits a greater performance gain with respect to the average accuracy, and almost all the categories on LIP exhibit better parsing predictions with the edge detection branch added. This indicates that boundary calibrations are especially effective for large-scale parsing items, which are typical and common in single human parsing datasets.

4.5 Evaluation of Rescoring Module

To further refine the parsing results, we introduce a parsing rescoring module at the end of the network. As the rescoring module is designed to exploit the merits of multibranch parsing results, we simply implement it with a convolutional layer without loss of generality. The ablation study of the kernel size is conducted on the LIP and CIHP datasets, and the comparison results are shown

Table 5. Comparison Results of the Parsing Rescoring Module with Respect to Different Kernel Sizes on the LIP Dataset

Method	pixel acc.	mean acc.	mIoU
Single human parsing on LIP			
w/o Rescore	87.99	67.43	55.38
w/ R_Conv1	87.97	67.43	55.52
w/ R_Conv7	88.07	67.74	55.61
w/ R_Conv15	88.04	67.89	55.70
w/ R_Conv31	**88.07**	**68.23**	**55.84**
Multiple human parsing on CIHP			
w/o Rescore	90.62	69.14	58.49
w/ R_Conv31	**90.74**	**69.68**	**59.07**

R_Conv* denotes the parsing rescoring module with kernel size of *. Note that large kernel sizes lead to explosively growing computational cost for vanilla convolutional layers. To be efficient, we implement it as the global convolution [32]. The best accuracies are indicated by bold fonts.

Table 6. Quantitative Comparisons of the Cost and Accuracies of the Baseline Model and Variants of the Proposed Method on the LIP Dataset

Method	#Params	GFLOPs	fps	mIoU
Baseline	66.7M	49.2	17.7	53.1
Ours	106.4M	162.6	14	55.8
Ours (Shared Params.)	74.1M	116.7	19.1	56.56

in Table 5. We can see that larger kernel sizes can bring consistent improvement with more computational cost. As the performance saturates when the kernel size reaches 31, we select this setting for the other experiments.

4.6 Network Analysis

To comprehensively analyze the proposed method, we show the numbers of parameters, running times, and accuracies of the baseline model and variants of our model in Table 6. We can see that the shared version of our method can outperform the unshared counterpart with smaller numbers of parameters and higher processing speed, which demonstrates the superior performance of the introduced multitask learning mechanism for human parsing. Specifically, compared with the baseline model, our method improves mIoU with 3.46 at the cost of an extra 7.4M parameters and 67.5 GFLOPs. In spite of the relatively high increase of GFLOPs, the runtime requirements can be satisfied in most of the real-world application scenarios.

4.7 Comparison with State-of-the-art Methods

In this section, we compare our method with the state of the art on three widely used benchmarks. Table 7 reports the comparison results on the LIP dataset, which is dedicated to single human parsing. We can see that the proposed method (denoted as "Ours*") can achieve superior performance across all three metrics even without the explicit hierarchical adjacent relationship constraints [12, 42]. In addition, we also conduct comparison experiments on the CIHP and PASCAL-Person-Part datasets, which are committed to multiple human parsing. The comparison results are summarized in Table 8 and Table 9. We can see from Table 8 that our method achieves

Table 7. Comparison Results of Our Method with State of the Art on the LIP Dataset [11]

Method	pixel acc.	mean acc.	mIoU
DeepLab (VGG-16)	82.66	51.64	41.64
DeepLab (ResNet-101)	84.09	55.62	44.80
JPPNet [11] (CVPR 17)	86.39	62.32	51.37
CE2P [35] (AAAI 19)	87.37	63.20	53.10
CNIF [40] (ICCV 19)	88.03	68.80	57.74
Grapy [12] (AAAI 20)	87.41	66.55	54.40
SCHP [18] (PAMI 20)	-	-	**59.36**
HTPR [42] (CVPR 20)	**89.05**	70.58	59.25
PCNet [51] (CVPR 20)	-	-	57.03
BGNet [50] (ECCV 20)	-	-	56.82
SNT [16] (ECCV 20)	88.10	70.41	54.86
Ours	88.30	68.50	56.56
Ours*	88.45	**71.12**	58.14

"Ours" and "Ours*" represent variants of the proposed method without and with the grid dividing technique, respectively, which is described in Section 3.2. The best accuracy is highlighted in bold font and the second and third best accuracies are underlined.

Table 8. Comparison Results of Our Method with State of the Art on the CIHP Dataset [10]

Method	mean acc.	mIoU
PGN [10] (ECCV 18)	64.22	55.80
DeepLab v3+[3] (ECCV 18)	65.06	57.13
Graphonomy (PASCAL) [9] (CVPR 19)	66.65	58.58
Grapy [12] (AAAI 20)	68.95	60.36
Grapy-ML [12] (AAAI 20)	68.97	60.60
PCNet [51] (CVPR 20)	67.05	61.05
HTPR [42] (CVPR 20)	72.67	60.60
Ours	72.36	61.91
Ours*	**72.98**	**62.61**

The best accuracy is highlighted in bold font and the second and third best accuracies are underlined.

Table 9. Comparison Results of Our Method with State of the Art on the PASCAL-Person-Part Dataset [4]

Method	mIoU
LIP [11] (CVPR 17)	59.36
RefineNet [25] (CVPR 17)	68.6
Bilinski et al. [1] (CVPR 18)	68.6
DeepLab v3+ [3] (ECCV 18)	67.84
PGN [10] (ECCV 18)	68.4
Graphonomy (CIHP) [9] (CVPR 19)	71.14
CNIF [40] (ICCV 19)	70.76
Grapy [12] (AAAI 20)	69.50
Grapy-ML [12] (AAAI 20)	71.65
SCHP [18] (PAMI 20)	71.46
HTPR [42] (CVPR 20)	73.12
PCNet [51] (CVPR 20)	**74.59**
BGNet [50] (ECCV 20)	74.42
SNT [16] (ECCV 20)	71.59
Ours	72.47
Ours*	73.46

The best accuracy is highlighted in bold font and the second and third best accuracies are underlined.

the best performance on CIHP with respect to all the metrics. It is worth noting that our method can consistently outperform hierarchical graph-based methods in multiple human parsing where crowd scenes are typically common, which verifies the effectiveness of the introduced multitask fusion mechanism. The comparison results in Table 9 demonstrate that the proposed method can be also effective on a small dataset.

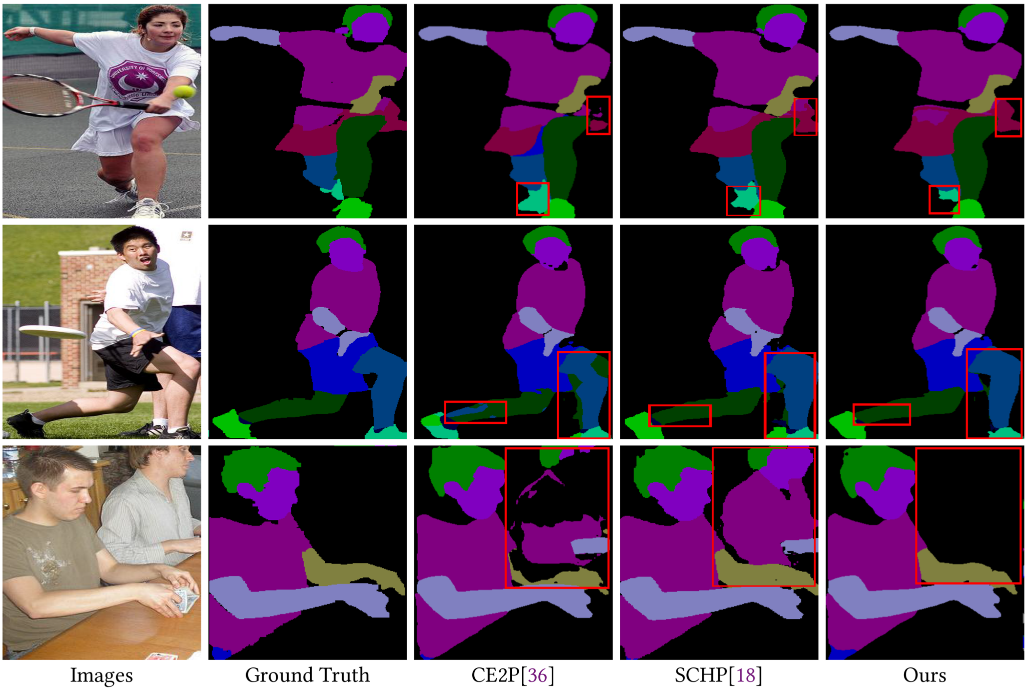


Fig. 6. Qualitative comparisons of the predicted human parsing results on the LIP dataset.

In addition to the quantitative comparison, we visualize the predicted human parsing results of the proposed method and state of the art on the LIP and CIHP datasets in Figure 6 and Figure 7. We can see that our method can make better judgments between foreground humans and background clutter and estimate clearer boundaries between parsing items. Notably, the three tasks have the potential to be conflicted as we can see in the first column of Figure 7 where "ours mask" missed the leftmost people. However, the mask prediction and the edge detection tasks provide supplements rather than compulsive guidance for human parsing. With the refinement of parsing predictions in the rescoring module, the possible misalignments within three tasks could be amended.

4.8 Failure Cases

We find that our model may fail in the following cases: (1) misprediction between categories with similar shapes (e.g., upper clothes and dress) when the human body is partly occluded or has an unusual posture, (2) overlapping with other shapes or textures in similar categories, and (3) inconsistent prediction within part of human items that is caused by noisy labels or largely changed image gradients. Corresponding examples are shown in Figure 8 in each column, respectively. To handle these problems, we may need to introduce extra guidance to compromise the local detail with the global context. Besides, a label refinement strategy should be promising to make a noisy-tolerant model in our further study.

5 CONCLUSION

In this article, motivated by several inherent problems of human parsing (e.g., label imbalance, scale variations, and nonrigid deformation), we tailor a mask-guided deformation adaptive

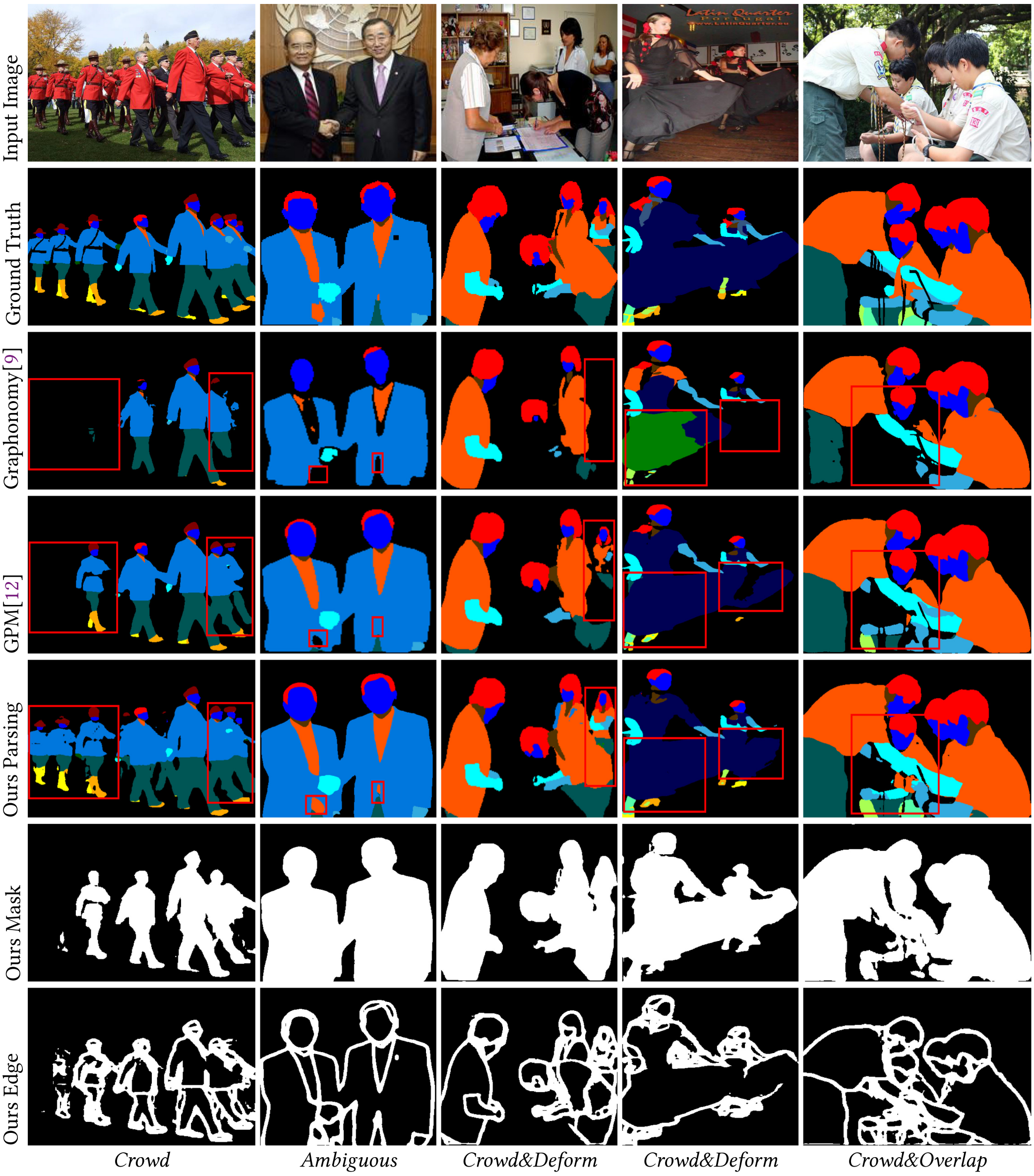


Fig. 7. Qualitative comparisons of the predicted human parsing results on the CIHP dataset [10]. For the convenience of discussion, we denote the challenging factors of each input image in the last row, where *Crowd*, *Ambiguous*, *Deform,* and *Overlap* represent crowd scenes, confusing foreground and background, more clothing deformation, and overlapped human body parts, respectively. It can be seen that our method can consistently alleviate the challenges thanks to the introduced human mask and edge detection, along with deformable convolution, and predict more accurate parsing results compared with competitors.

network to resolve these problems. Specifically, we introduce the human mask and edge detection branch as the auxiliary task with deformable convolution for learning rich and diverse human parsing feature representations. This not only relieves the inferior predictions caused by the imbalance of background/foreground separations but also readjusts the convolutional sampling offsets

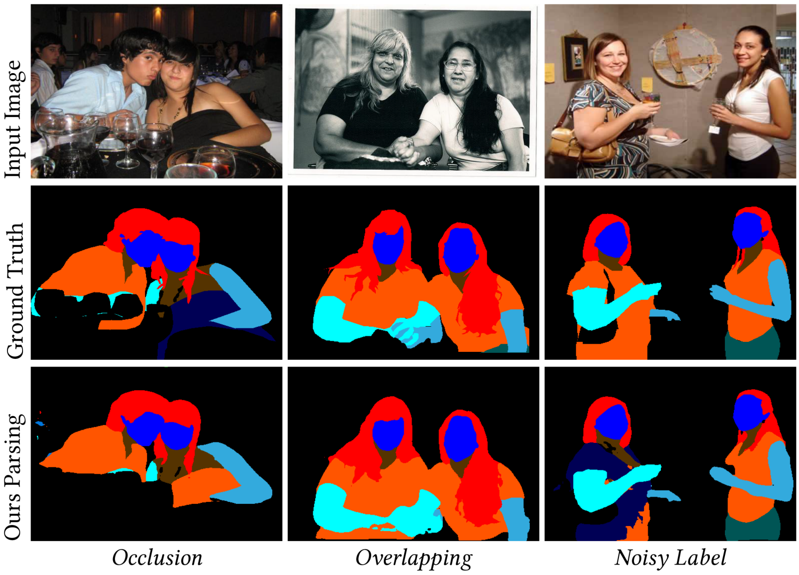


Fig. 8. Failure cases. The first column shows that the occluded dress of the right people is mispredicted as upper clothes. The second column shows that the cross-overlapped left arm of the left people is mispredicted as right arm. The last column shows an inconsistent prediction case where the dress of the left people was wrongly labeled as upper clothes, which is partly mispredicted as upper clothes by our method.

thanks to deformable convolution. Comprehensive experiments show that the proposed method can outperform state-of-the-art methods on both the single human and multiple human parsing datasets.

REFERENCES

[1] Piotr Bilinski and Victor Prisacariu. 2018. Dense decoder shortcut connections for single-pass semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 6596–6605.

[2] Liang-Chieh Chen, Jonathan T. Barron, George Papandreou, Kevin Murphy, and Alan L. Yuille. 2016. Semantic image segmentation with task-specific edge detection using CNNs and a discriminatively trained domain transform. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 4545–4554.

[3] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. 2018. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *Proceedings of the European Conference on Computer Vision*.

[4] Xianjie Chen, Roozbeh Mottaghi, Xiaobai Liu, Sanja Fidler, Raquel Urtasun, and Alan Yuille. 2014. Detect what you can: Detecting and representing objects using holistic models and body parts. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 1979–1986.

[5] J. Dai, H. Qi, Y. Xiong, Y. Li, G. Zhang, H. Hu, and Y. Wei. 2017. Deformable convolutional networks. In *Proceedings of IEEE International Conference on Computer Vision*. IEEE, 764–773.

[6] Hao-Shu Fang, Guansong Lu, Xiaolin Fang, Jianwen Xie, Yu-Wing Tai, and Cewu Lu. 2018. Weakly and semi supervised human body part parsing via pose-guided knowledge transfer. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 70–78.

[7] Pedro F. Felzenszwalb, Ross B. Girshick, David A. McAllester, and Deva Ramanan. 2009. Object detection with discriminatively trained part based models. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 32 (2009), 1627–1645.

[8] Jun Fu, Jing Liu, Haijie Tian, Yong Li, Yongjun Bao, Zhiwei Fang, and Hanqing Lu. 2019. Dual attention network for scene segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 3141–3149.

[9] Ke Gong, Yiming Gao, Xiaodan Liang, Xiaohui Shen, Meng Wang, and Liang Lin. 2019. Graphonomy: Universal human parsing via graph transfer learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 7442–7451.

[10] Ke Gong, Xiaodan Liang, Yicheng Li, Yimin Chen, Ming Yang, and Liang Lin. 2018. Instance-level human parsing via part grouping network. In *Proceedings of the European Conference on Computer Vision*. Springer, Cham, 805–822.

[11] Ke Gong, Xiaodan Liang, Dongyu Zhang, Xiaohui Shen, and Liang Lin. 2017. Look into person: Self-supervised structure-sensitive learning and a new benchmark for human parsing. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 6757–6765.

[12] Haoyu He, Jing Zhang, Qiming Zhang, and Dacheng Tao. 2020. Grapy-ML: Graph pyramid mutual learning for cross-dataset human parsing. In *Proceedings of the AAAI Conference on Artificial Intelligence*. AAAI.

[13] K. He, X. Zhang, S. Ren, and J. Sun. 2016. Deep residual learning for image recognition. In *CVPR*. IEEE, 770–778.

[14] Jie Hu, Li Shen, and Gang Sun. 2018. Squeeze-and-excitation networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 7132–7141.

[15] Max Jaderberg, Karen Simonyan, Andrew Zisserman, and Koray Kavukcuoglu. 2015. Spatial transformer networks. In *NeurIPS*. Curran Associates, Inc., Montreal, Quebec, Canada.

[16] Ruyi Ji, Dawei Du, Libo Zhang, Longyin Wen, Yanjun Wu, Chen Zhao, Feiyue Huang, and Siwei Lyu. 2020. Learning semantic neural tree for human parsing. In *ECCV*, Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm (Eds.).

[17] Mahdi M. Kalayeh, Emrah Basaran, Muhittin Gokmen, Mustafa E. Kamasak, and Mubarak Shah. 2018. Human semantic parsing for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 1062–1071.

[18] P. Li, Y. Xu, Y. Wei, and Y. Yang. 2020. Self-correction for human parsing. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (Early Access) (2020), 1–1.

[19] T. Li, Z. Liang, S. Zhao, J. Gong, and J. Shen. 2020. Self-learning with rectification strategy for human parsing. In *CVPR*. 9260–9269.

[20] Yanwei Li, Xinze Chen, Zheng Zhu, Lingxi Xie, Guan Huang, Dalong Du, and Xingang Wang. 2019. Attention-guided unified network for panoptic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 7019–7028.

[21] X. Liang, K. Gong, X. Shen, and L. Lin. 2019. Look into person: Joint body parsing pose estimation network and a new benchmark. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 41 (2019), 871–885.

[22] X. Liang, L. Lin, Y. Wei, X. Shen, J. Yang, and S. Yan. 2018. Proposal-free network for instance-level object segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 40 (2018), 2978–2991.

[23] X. Liang, L. Lin, W. Yang, P. Luo, J. Huang, and S. Yan. 2016. Clothes co-parsing via joint image segmentation and labeling with application to clothing retrieval. *IEEE Transactions on Multimedia* 18 (2016), 1175–1186.

[24] X. Liang, S. Liu, X. Shen, J. Yang, L. Liu, J. Dong, L. Lin, and S. Yan. 2015. Deep human parsing with active template regression. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 37 (2015), 2402–2414.

[25] Guosheng Lin, Fayao Liu, Anton Milan, Chunhua Shen, and Ian Reid. 2019. RefineNet: Multi-path refinement networks for dense prediction. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 42 (2019), 1228–1242.

[26] Yutian Lin, Liang Zheng, Zhedong Zheng, Yu Wu, and Yang Yang. 2019. Improving person re-identification by attribute and identity learning. *Pattern Recognition* 95 (2019), 151–161.

[27] S. Liu, J. Feng, C. Domokos, H. Xu, J. Huang, Z. Hu, and S. Yan. 2014. Fashion parsing with weak color-category labels. *IEEE Transactions on Multimedia* 16 (2014), 253–265.

[28] S. Liu, X. Liang, L. Liu, X. Shen, J. Yang, C. Xu, L. Lin, Xiaochun Cao, and S. Yan. 2015. Matching-CNN meets KNN: Quasi-parametric human parsing. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 1419–1427.

[29] Ping Luo, Xiaogang Wang, and Xiaoou Tang. 2013. Pedestrian parsing via deep decompositional network. In *Proceedings of IEEE International Conference on Computer Vision*. IEEE, 2380–7504.

[30] Yawei Luo, Zhedong Zheng, Liang Zheng, Tao Guan, Junqing Yu, and Yi Yang. 2018. Macro-micro adversarial network for human parsing. In *Proceedings of the European Conference on Computer Vision*. Springer, Cham, Munich, Germany, 424–440.

[31] Xuecheng Nie, Jiashi Feng, and Shuicheng Yan. 2018. Mutual learning to adapt for joint human parsing and pose estimation. In *Proceedings of the European Conference on Computer Vision*. Springer, Cham, Munich, Germany, 519–534.

[32] Chao Peng, Xiangyu Zhang, Gang Yu, Guiming Luo, and Jian Sun. 2017. Large kernel matters–improve semantic segmentation by global convolutional network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 4353–4361.

[33] Xuebin Qin, Zichen Zhang, Chenyang Huang, Chao Gao, Masood Dehghan, and Martin Jagersand. 2019. BASNet: Boundary-Aware salient object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 7471–7481.

[34] Rodolfo Quispe and Helio Pedrini. 2019. Enhanced person re-identification based on saliency and semantic parsing with deep neural network models. *Image and Vision Computing* 92 (2019), 103809.

[35] Tao Ruan, Ting Liu, Zilong Huang, Yunchao Wei, Shikui Wei, and Yao Zhao. 2019. Devil in the details: Towards accurate single and multiple human parsing. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 33. 4814–4821.

[36] Tao Ruan, Ting Liu, Zilong Huang, Yunchao Wei, Shikui Wei, Yao Zhao, and Thomas Huang. 2019. Devil in the details: Towards accurate single and multiple human parsing. In *Proceedings of the AAAI Conference on Artificial Intelligence*. AAAI, 4814–4821.

[37] A. Shahroudy, T. Ng, Q. Yang, and G. Wang. 2016. Multimodal multipart learning for action recognition in depth videos. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 38 (2016), 2123–2129.

[38] Ilya Sutskever, James Martens, George Dahl, and Geoffrey Hinton. 2013. On the importance of initialization and momentum in deep learning. In *Proceedings of International Conference on Machine Learning*. PMLR, Atlanta, Georgia, 1139–1147.

[39] Towaki Takikawa, David Acuna, Varun Jampani, and Sanja Fidler. 2019. Gated-SCNN: Gated shape CNNs for semantic segmentation. In *Proceedings of IEEE International Conference on Computer Vision*. IEEE, 5228–5237.

[40] Wenguan Wang, Zhijie Zhang, Siyuan Qi, Jianbing Shen, Yanwei Pang, and Ling Shao. 2019. Learning compositional neural information fusion for human parsing. In *Proceedings of IEEE International Conference on Computer Vision*. IEEE, 5702–5712.

[41] W. Wang, T. Zhou, S. Qi, J. Shen, and S. C. Zhu. 2021. Hierarchical human semantic parsing with comprehensive part-relation modeling. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (Early Access) (2021), 1–1. https://doi.org/10.1109/TPAMI.2021.3055780

[42] W. Wang, H. Zhu, J. Dai, Y. Pang, J. Shen, and L. Shao. 2020. Hierarchical human parsing with typed part-relation reasoning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 8926–8936.

[43] Yang Wang, Duan Tran, Zicheng Liao, and David A. Forsyth. 2012. Discriminative hierarchical part-based models for human parsing and action recognition. *Journal of Machine Learning Research* 13 (2012), 3075–3102.

[44] Y. Wu, Y. Lin, X. Dong, Y. Yan, W. Bian, and Y. Yang. 2019. Progressive learning for person re-identification with one example. *IEEE Transactions on Image Processing* 28 (2019), 2872–2881.

[45] Fangting Xia, Jun Zhu, Peng Wang, and Alan L. Yuille. 2016. Pose-Guided human parsing by an and/or graph using pose-context features. In *Proceedings of the AAAI Conference on Artificial Intelligence*. AAAI, 3632–3640.

[46] Saining Xie and Zhuowen Tu. 2017. Holistically-nested edge detection. *International Journal of Computer Vision* 125 (2017), 3–18.

[47] Yuwen Xiong, Renjie Liao, Hengshuang Zhao, Rui Hu, Min Bai, Ersin Yumer, and Raquel Urtasun. 2019. UPSNet: A unified panoptic segmentation network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 8810–8818.

[48] K. Yamaguchi, M. H. Kiapour, L. E. Ortiz, and T. L. Berg. 2015. Retrieving similar styles to parse clothing. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 37 (2015), 1028–1040.

[49] Zhiding Yu, Chen Feng, Ming-Yu Liu, and Srikumar Ramalingam. 2017. CASENet: Deep category-aware semantic edge detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 1761–1770.

[50] Xiaomei Zhang, Y. Chen, B. Zhu, Jinqiao Wang, and Ming Tang. 2020. Blended grammar network for human parsing. In *Proceedings of the European Conference on Computer Vision*.

[51] X. Zhang, Y. Chen, B. Zhu, J. Wang, and M. Tang. 2020. Part-aware context network for human parsing. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 8968–8977.

[52] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia. 2017. Pyramid scene parsing network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 6230–6239.

[53] J. Zhao, J. Li, X. Nie, F. Zhao, Y. Chen, Z. Wang, J. Feng, and S. Yan. 2017. Self-Supervised neural aggregation networks for human parsing. In *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW'17)*. IEEE.

[54] Ting Zhao and Xiangqian Wu. 2019. Pyramid feature attention network for saliency detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 3080–3089.

[55] Bingke Zhu, Yingying Chen, Ming Tang, and Jinqiao Wang. 2018. Progressive cognitive human parsing. In *Proceedings of the AAAI Conference on Artificial Intelligence*. AAAI.

[56] Xizhou Zhu, Han Hu, Stephen Lin, and Jifeng Dai. 2019. Deformable ConvNets V2: More deformable, better results. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 9300–9308.

Received August 2020; revised March 2021; accepted May 2021