Correlating Edge, Pose with Parsing

Ziwei Zhang[1], Chi Su[2*], Liang Zheng[3], Xiaodong Xie[1]
[1]School of Electronics Engineering and Computer Science, Peking University
[2]Kingsoft Cloud, [3]Australian National University
{ziwei.zh,donxie}@pku.edu.cn, suchi@kingsoft.com, liang.zheng@anu.edu.au

Abstract

According to existing studies, human body edge and pose are two beneficial factors to human parsing. The effectiveness of each of the high-level features (edge and pose) is confirmed through the concatenation of their features with the parsing features. Driven by the insights, this paper studies how human semantic boundaries and keypoint locations can jointly improve human parsing. Compared with the existing practice of feature concatenation, we find that uncovering the correlation among the three factors is a superior way of leveraging the pivotal contextual cues provided by edges and poses. To capture such correlations, we propose a Correlation Parsing Machine (CorrPM) employing a heterogeneous non-local block to discover the spatial affinity among feature maps from the edge, pose and parsing. The proposed CorrPM allows us to report new state-of-the-art accuracy on three human parsing datasets. Importantly, comparative studies confirm the advantages of feature correlation over the concatenation.

1. Introduction

This paper studies human parsing, aiming to partition a human image into semantic regions including body parts and clothes. This problem is challenging due to the complicated textures and styles of clothes, the deformable human body, the scale diversity of different categories, *etc*. As such, directly applying general semantic segmentation methods to human parsing may lead to unsatisfying results, which are reflected in two aspects. First, the boundaries between adjacent parts may be inaccurately located. The system might get confused with pixels along the boundaries, especially when the neighboring parts have similar appearance. Second, semantics of segmented parts may be inconsistent with human body structure, if we don not consider the affinity among different parts. This leads to mislabeling or missing predictions when context clues are not obvious.

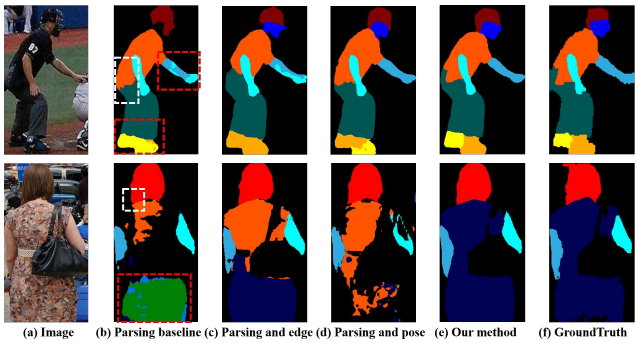

Figure 1. Illustration of parsing errors and our motivation. (a) Given images. (b) Results of parsing baseline [4]. (c) Fusion of parsing and human body edge features. (d) Fusion of parsing and the human keypoint features. (e) Results of our method. (f) Groundtruth. From (b), we observe parsing errors happen due to boundary ambiguity (white box) and body structure inconsistency (red box), respectively. The fusion of boundary features (c) or keypoint features (d) may mitigate one of the two errors. The two types of errors are obviously mitigated in (e) because we take the advantage of both boundary and keypoints by learning their correlation with parsing. By comparison, the proposed strategy is superior to concatenation or post processing as commonly done.

Edge detection and pose estimation can potentially address the above two problems. For the first problem, *i.e.,* boundary confusions, human body edge detection is beneficial to distinguish two adjacent categories [3, 33]. For the second problem, *i.e.,* semantics inconsistency, pose estimation provides keypoints to enforce the parsing results to be semantically consistent with human body structure [38, 28, 34]. Therefore, current research [3, 33, 37, 29] identifies human edge and pose as complementary cues to improve parsing performance. As shown in Fig. 1(b), when directly using generic segmentation methods for human parsing, some pixels of upper clothes are predicted as pants: the network incorrectly locates the edges between the two categories. Moreover, due to the lack of human semantic constraints, the left and right arms, left and right shoes are incorrectly identified. In Fig. 1(c), after adding edge

*Corresponding Author.
Code is available at: https://github.com/ziwei-zh/CorrPM.

information to parsing, we observe that the boundary pixels are accurately located. Further, when utilizing body part cues provided by pose features in Fig. 1(d), the mistaken prediction of the left arm no longer exists, and the left shoe is clearly distinguished from the right shoe.

In spite of the improvements so far, existing research utilising edge/pose to improve parsing has not leveraged them to the full potential. Usually a single factor, *i.e.* either pose or edge, is used, which might be beneficial to handle *a single problem* mentioned above. In addition, existing methods typically perform feature concatenation or post processing for parsing refinement. We point out that this practice might be inferior. As shown in Fig. 1(c) and (d), when only a single factor is concerned for parsing system, there still remain blurs and holes in arm and dress area, and left/right shoes are inexactly predicted. Therefore, simple fusion or post processing may not be enough to process fine regions, such as edges of different parts. To address this problem, we explore the correlations among edge and pose and find that it is preferable that edge, pose and parsing are simultaneously integrated.

In this paper, we propose a Correlation Parsing Machine (CorrPM) to take advantage of both human semantic edge and pose features to benefit human parsing. Contrary to performing feature concatenation or post processing, we learn the correlation among the three tasks. The CorrPM has three encoders and is featured by a heterogeneous non-local (HNL) module. The encoders calculate vector representations of the human edge, pose and semantics, respectively. HNL mixes the three features into a hybrid representation and explores the spatial affinity between this hybrid feature and the parsing feature map at all positions. As such, our method can effectively perceive the human edges and maintains the integrity of a semantic region, addressing the inaccurate boundary localization problem. Meanwhile, by perceiving the body keypoints, our method improves the consistency of the body part geometry. For example, as shown in Fig. 1 (e), our method corrects the mislabeling of arm region and correctly segments the boundary between upper clothes and pants, and between dresses and arms.

To summarize, our contribution is three-fold. 1) We propose to use a Heterogeneous Non-Local (HNL) structure to capture the correlations among three closely related factors. 2) We show that human edge and pose, when both integrated in the Correlation Parsing Machine (CorrPM), bring significant improvement to parsing task. 3) Using simple edge detection and pose estimation models, we report very competitive parsing accuracy on three human parsing datasets.

2. Related Work

Semantic Segmentation. Human parsing is a fine-grained semantic segmentation, which performs per-pixel prediction on all objects. Due to its great prospects in application, semantic segmentation has gained much importance in the past few years. FCN [25, 5, 42] performs well on this task which applies fully convolution on the whole image to produce labels of every pixel. Inspired by this, many researchers [31, 1, 32] start to leverage the encoder-decoder structure which extracts features by downsampling and then use upsampling to recover them to the original resolution. Aiming to enlarge the receptive field, another structure, DeepLab [4], designs atrous convolution kernels to force the network to perceive larger area and reduce the prediction errors. Zhao *et al.* [41] propose a pyramid scene parsing network aggregating multi-scale object clues to make the segmentation more precise. In [36], Xia *et al.* propose the "Auto-zoom Nets" to automatically "zoom" the objects and parts which have diverse scales.

Human Parsing. Following main approaches in semantic segmentation, early researches in human parsing contribute towards this topic mostly by hand-crafted features and post-processed by Conditional Random Field (CRF) [11, 23]. Dong *et al.* [9] use a variety of parselets assembled by "And-Or" sub-trees to jointly parse human body labels and keypoint locations. With the development of convolutional neural network (CNN), especially after the ResNet [17] is proposed, many deep learning approaches have achieved much progress in this area. In [22], Liang *et al.* propose a Co-CNN framework capturing cross-layer local and global context information to boost the parsing performance. Gong *et al.* [15] introduce a new large-scale benchmark LIP and a novel self-supervised structure-sensitive learning method. In [20], Li *et al.* tackle the human parsing problem by generating global parsing maps for person in a bottom-up way.

Utilizing edge or pose for parsing. Aiming to get more accurate predictions in human parsing task, recent works [10, 29, 14, 15, 37, 13, 7, 19, 30] utilize edge or pose information as a guidance. Chen *et al.* [3] propose an edge-aware filtering method to capture accurate semantic contours between two adjacent parts. Ruan *et al.* [33] fuse the edge map with parsing feature which can reserve the boundary of person parts to benefit the human parsing. Gong *et al.* [14] conduct both semantic part parsing and edge detection in the way of sharing intermediate representation of both features. Xia *et al.* [37] train two FCNs to predict poses and parts separately and then fuse them through a fully-connected conditional random field (FCRF) as a refinement. Nie *et al.* [29] observe that pose and parsing can simultaneously boost the performance of each other by training two parallel models and adapt the mutual parameters. Despite the improvement, the existing methods simply perform feature concatenation or pose-processing to refine parsing results, which is inferior to guide parsing model to learn contextual cues. Our framework simultaneously integrates edge, pose and parsing representation and effectively

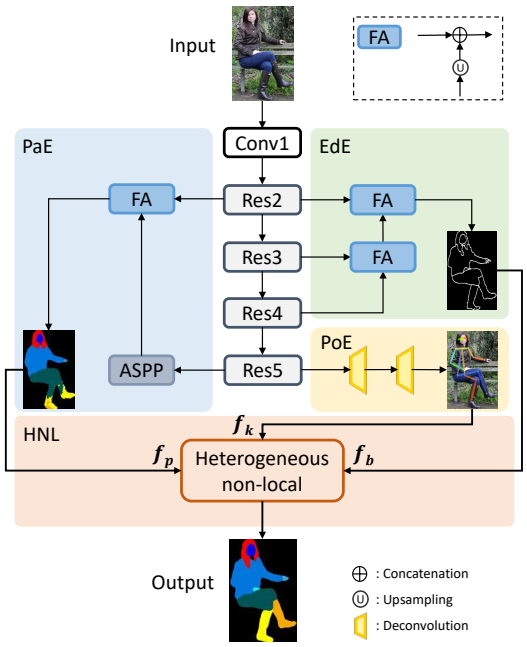

Figure 2. Overview of the proposed network. PaE: parsing encoder. EdE: edge encoder. PoE: pose encoder. HNL: heterogeneous non-local module. FA: feature aggregation. f_*: parsing/edge/pose features. After extracted by three encoders, parsing, pose and edge features are fed into HNL to explore their correlation to benefit human parsing task.

exploits the correlation among these three representation.

Non-local Network. Human parsing is closely complementary to semantic edge information and human pose information. And the relationship among them is exploited and employed by HNL which is modified from non-local network. Originating from non-local means algorithm [2] , the non-local network is leveraged in many approaches to capture long-range dependencies [43, 40]. Wang *et al.*[35] propose the non-local block as a weighted summation of relationships of every position and show good performance in video classification. Even though non-local network has been a great success in many tasks , existing methods seek the relationship with the feature itself. Different from the existing self-attention mechanism, the proposed heterogeneous non-local module aggregates parsing, edge and pose factors together and learns the correlation of parsing with the other two features.

3. The Proposed Approach

As illustrated in Fig. 2, the proposed Correlation Parsing Machine (CorrPM) leverages human body keypoint and semantic boundary information to benefit human parsing. We firstly introduce the overall formulation of our framework in Section 3.1. The three feature encoders are represented in Section 3.2 and we propose a heterogeneous non-local module (HNL) to correlate the three factors in Section 3.3. Then, Section 3.4 explains the difference between the proposed HNL and the traditional non-local networks. And the overall training objective is illustrated in Section 3.5.

3.1. Formulation

Given an input image $I \in \mathbb{R}^{3\times M\times N}$ of size $M \times N$, our task is to predict the label of every pixel and generate a segmentation mask $P \in \mathbb{R}^{M\times N}$ leveraging three kinds of information: human body part category $\mathcal{P} \in \{0, 1, ..., Q\}^{M\times N}$, semantic boundary $\mathcal{B} \in \{0, 1\}^{M\times N}$ and human body keypoint location $\mathcal{K} = \{(x_i, y_i)\}_{i=1}^{J}$. J and Q are the number of body joints and part categories. (x_i, y_i) are the coordinates of the point i, and the pixels that belong to boundaries are labeled as 1 otherwise 0. We aim to design a unified framework that jointly utilizes these three factors and uncovers the correlation among them to better leverage the pivotal contextual cues.

3.2. Feature Encoding

Human parsing, pose estimation, edge detection are complementary and closely related, hence, their features can be learned by a shared base model Θ, *e.g.*, ResNet101 [17]. The feature at the lower stage of the base model retains high resolution structure and is fed to edge encoder to capture the object edge boundaries f_b. And the higher-stage feature keeps rich semantic information which is further used as parsing feature f_p and keypoint feature f_k.

Parsing Encoder. We adopt a parsing pipeline to predict a coarse segmentation map firstly. Context information is leveraged in many previous work [41, 4] in semantic segmentation and it is also essential in human parsing. Given the parsing feature of the base model Θ, we observe that merely performing dense pixel-wise prediction on it will cause mislabeling. Therefore, we add the Atrous Spatial Pyramid Pooling (ASPP) [4] to enlarge the receptive fields and get more useful context cues.

Meanwhile, some objects in human parsing have quite low resolution, *e.g.*, sunglasses and socks, so the details might be lost in the process of downsampling. We employ the feature of *Res2* of base model and upsample the output of ASPP module to the same scale as *Res2* and concatenate them as f_p. After extracted from the parsing encoder, the feature f_p obtains a coarse semantic representation and will be further fed into the heterogeneous non-local module to obtain pose and edge guidance.

Pose Encoder. In order to get human body structure cues, we design a pose encoder to get joint locations. Many existing approaches [37, 29, 28] in pose estimation adopt complicated CNNs to get more accurate keypoint locations. For instance, Hourglass [28] performs repeated downsampling and upsampling procedure to capture multi-scale keypoint information. Different from them, we only deploy two

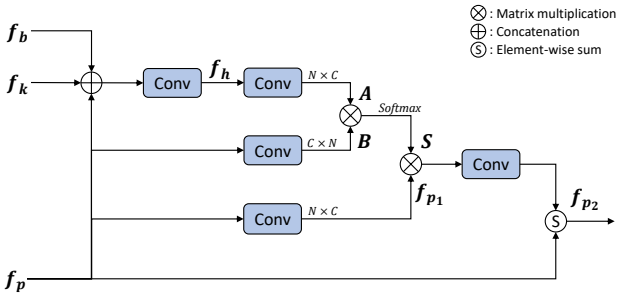

Figure 3. Structure of the heterogeneous non-local (HNL) module. It aggregates parsing, edge and pose feature into a hybrid feature f_h and calculates the correlation between f_h and f_p.

transposed convolution layers [38] to extract human keypoint structure, since pose estimation task can also get benefits from the parsing task. As a result, the shared feature is upsampled by 4 times generating the pose feature f_k. It is the same scale as the parsing feature f_p.

After the pose representation f_k is captured, we regress the heatmap from it. Following [34], we apply 2D Gaussian filter centered on each annotated keypoint coordinate with standard deviation of 7 pixels, and generate the ground truth heatmap as the supervision of pose encoder.

Edge Encoder. In human parsing task, semantic boundary ambiguity remains to be solved. The border pixels of two adjacent semantic parts may be inaccurately predicted, particularly when they have similar appearances. Hence, we propose an edge encoder to learn feature f_b with boundary consciousness. It is observed that lower stages in neural network maintain high resolution and higher-stage feature obtains detailed semantic information. As shown in Fig. 2, we leverage the features of *Res2*, *Res3* and *Res4* which retain both large spatial details and semantic consistency. The feature maps are upsampled to the same size as *Res2* by linear interpolation. Then, they are concatenated and fed into a 1×1 convolution layer to generate the edge feature map f_b. The edge encoder is supervised by the edge information between two adjacent categories and the feature will be further fed into the heterogeneous non-local correlation block.

3.3. Heterogeneous Non-Local

Many existing researches prove that either edge or pose is a beneficial factor to parsing task. However, the fusion strategy they employ cannot fully leverage the two factors as discussed above. Recently, the correlation module is used to capture the long-range contextual information by self-attention [12, 18] or explore the relationship between the two features [43]. However, if we follow this operation, the correlation computation cost is high ($O(n^2)$, n is the number of feature maps) and the overall model is hard to converge. Therefore, we propose a Heterogeneous Non-Local (HNL) block to fully leverage the contextual cues provided by boundaries and poses, which we believe is more effective and more efficient.

As shown in Fig. 3, we first aggregate the three factors by concatenating them in the channel dimension, and then a convolution layer parameterized by W_a is conducted to transform it into a hybrid feature f_h, whose dimension is the same as that of $f_p \in \mathbb{R}^{C \times H \times W}$:

$$f_h = W_a(f_p \oplus f_b \oplus f_k), \tag{1}$$

where $\oplus$ means concatenation.

We exchange the self-attention in the standard non-local block with correlation between the hybrid feature f_h and parsing feature f_p. First, f_h and f_p are fed into two convolution layers to generate two new features A and B, then we reshape them into matrixes with size $N \times C$ and $C \times N$, respectively, where $N = H \times W$ denotes the total number of pixels per channel. We compute the relationship map $\boldsymbol{S} \in \mathbb{R}^{N \times N}$ by a matrix product of A and B, and normalize the relation map by a softmax operation.

$$\boldsymbol{S} = softmax(A \cdot B) \tag{2}$$

where a point (i, j) in $\boldsymbol{S}$ measures the relation affinity between the i^{th} pixel in hybrid feature f_h and j^{th} pixel in parsing feature f_p.

Then we feed the parsing feature f_p into another convolution layer to generate $f_{p_1} \in \mathbb{R}^{C \times H \times W}$ and reshape it to $\mathbb{R}^{N \times C}$, which is multiplied by $\boldsymbol{S}$ to integrate the pixel correlation cues into parsing features. The resulting feature is fed into the final convolution layer parameterized by W_b and added back to f_p element-wise to get the final parsing feature f_{p_2}. The overall procedure can be formulated as:

$$f_{p_2} = W_b(\boldsymbol{S} \cdot f_{p_1}) + f_p, \tag{3}$$

where W_b is initialized as 0. In this way, the hybrid representation effectively aggregates parsing, edge and pose information together. And the refined parsing feature f_{p_2} in Equation 3 is a weighted summation of every position in the hybrid feature and the parsing feature. Therefore, it obtains edge information between two bordered parts and retains semantic consistency with human body, thus getting more reasonable parsing results.

3.4. Discussions

The heterogeneous non-local block is an extension of non-local neural network [35]. However, different from the traditional non-local operation which only computes the relationship of one feature as a mechanism of self-attention, the proposed network has three advantages. First, it integrates human parsing, pose estimation and edge detection tasks into a unified model, and the correlation is calculated among three different feature representations. Second, HNL does not add much computation complexity compared with traditional non-local structure while maintains

very competitive accuracy. Finally, for other tasks which are also related to human parsing, it has potential to integrate it into the hybrid representation and model the relationship among them with only little computation complexity increased (brought by the corresponding encoder).

3.5. Training objectives

In addition to the parsing supervision, human keypoint location and semantic edge information are utilized to train the whole model. The total training objective is:

$$L = L_{p_2} + L_p + \alpha L_b + \beta L_k, \tag{4}$$

L_{p_2} or L_p is the loss between the parsing result f_{p_2} or f_p and the parsing annotations; L_b denotes the loss between the predicted edge map f_b and the edge annotation; L_k is the loss between the body joints prediction f_k and the ground truth coordinates. It is worth noting that the edge annotation is obtained by finding the borders of the mask between two different semantic parts, which needs no additional annotations. Cross Entropy loss is adopted as L_{p_2}, L_p and L_b, and Mean Square Error loss is used for L_k. The whole framework is trained end-to-end.

4. Experiments

4.1. Experimental Settings

Datasets and metrics. We evaluate the performance of the proposed method on three human parsing datasets:

LIP [15] is a large-scale benchmark dataset focusing on semantic understanding of human body parts and clothes labels. It contains coordinates of 16 body keypoints and pixel-level annotations of 20 semantic human parts (including one background label). There are totally 50,462 images which are further split into train/val/test sets containing 30,462/10,000/10,000 images, respectively.

ATR [22] contains 18 categories of human part labels including *face, sunglasses, hat, scarf, hair, upper-clothes, left/right arm, belt, pants, left/right leg, skirt, left/right shoe, bag, dress* and *background*. Following [22], we use 16,000 images for training, 1,000 for testing and 700 for validation.

CIHP [14] provides 38,280 images with 20 categories. It contains 28,280 training, 5,000 validation and 5,000 test images. On account of no human pose annotations in ATR and CIHP, we utilize the pose estimator [38] trained on COCO [24] to obtain human body keypoint locations as ground truth. During training, we first utilize Mask R-CNN[16] to generate the mask of every person, and apply it on multi-person images to generate single person images. We obtain 93,213 training images in total. During inference, single person is segmented from background in the same way as training and we conduct parsing with the proposed network and finally merge them into the original image.

Method	EA	BI	Fusion Strategy	Accuracy
[33]	✓		Feature concatenation	++
[14]	✓		Feature concatenation	++
[29]		✓	Parameters mutual learning	++
[15]		✓	Loss constraint	+
[37]		✓	Post processing	+
Ours	✓	✓	Correlation	+++

Table 1. Comparison of different fusion methods. EA represents edge ambiguity and BI represents boundary inconsistency. Existing methods use either edge or pose to solve a single problem in parsing. Different from them, we aggregate parsing, edge and pose feature and explore the correlation among them which shows the superiority on Accuracy.

We report Accuracy, mIoU, Precision, Recall and F-1 score to evaluate the parsing performance on the datasets.

Training Details. We train CorrPM from scratch for 150 epochs, and adopt ResNet101 [17] pre-trained on ImageNet [8] as the base model Θ. During training, the 384×384 input images are randomly rotated (from -60° to 60°), flipped and resized (from 0.75 to 1.25). f_p, f_k and f_b are in the same size of $C \times H \times W$, where $C = 512$ and $H = W = 96$. We use SGD as the optimizer and the learning rate is initially set as 1e-3. Following previous works [44], we employ the "poly" learning rate policy, and the learning rate is multiplied by $(1 - \frac{iter}{total\_iter})^{0.9}$. We set the momentum to 0.9 and weight decay to 5e-4. The edge loss weight α and pose loss weight β are 2 and 70.

Testing phase. During inference, the outputs of pose and edge branches are ignored, and f_{p_2} is employed to predict the final parsing mask P. The inference procedure is executed on a 12GB TITAN V for a fair speed comparison with other methods. Our model does not add too much complexity compared with direct concatenation, because the base model (ResNet-101) consumes a majority of computations. CorrPM achieves a speed of 11 fps which is faster than Attention+SSL [15] (2 fps) and MuLA [29] (5 fps).

4.2. Comparison with related methods

4.2.1 Fusion strategy comparison

Tab. 1 lists some existing researches that utilize pose or edge information to assist human parsing task. For edge ambiguity issue, [14] and [33] extract edge feature and concatenate it with parsing feature to perceive useful cues of part boundaries. But this fusion strategy is not able to sufficiently obtain the semantic boundary completeness. Aiming to solve body inconsistency problem, [29] conducts two parallel human pose estimation and human parsing networks and mutually learns the parameters. However, the training process is somewhat complicated. Meanwhile, [37] adopts FCRF as a way of post-processing and [15] adds a joint loss utilizing the pose information to constrain part segments. Above fusion methods only employ a single factor and merely handle a single problem. In comparison,

Method	Pixel Acc.	Mean Acc.	mIoU
DeepLabV2 [4]	82.66	51.64	41.64
Attention [5]	83.43	54.39	42.92
Attention+SSL [15]	84.36	54.94	44.73
SS-NAN [42]	87.59	56.03	47.92
MuLA(Hourglass) [29]	**88.50**	60.50	49.30
JPPNet [21]	86.39	62.32	51.37
CE2P [33]	87.37	63.20	53.10
Ours†	87.36	66.37	54.43
Ours	87.68	**67.21**	**55.33**

Table 2. Comparison of different methods on the validation set of the LIP dataset. † means removing L_p in Equation 4.

Method	Acc	F.g.Acc	Pre	Rec	F-1 score
DeepLabV2 [4]	94.42	82.93	69.24	78.48	73.53
Attention [5]	95.41	85.71	81.30	73.55	77.23
CoCNN [22]	96.02	83.57	84.59	77.66	80.14
TGPNet [26]	96.45	87.91	83.36	80.22	81.76
Ours	**97.12**	**90.40**	**89.18**	**83.93**	**86.12**

Table 3. Comparison of Accuracy, Foreground Accuracy, Precision, Recall and F-1 score on the ATR test set.

our CorrPM combines parsing with pose and edge information, and the experiment also shows exploring the correlation among the three factors is a superior feature fusion strategy to other recent methods.

4.2.2 Performance on single-person datasets

LIP. We show the performance comparison of the proposed model and the other methods on LIP validation set. As shown in Tab. 2, the proposed CorrPM achieves the best performance of 55.33% in terms of mIoU and significantly outperforms other methods. Specifically, JPPNet and MuLA add pose supervision as a constraint of human parsing. CE2P adds edge information to refine parsing results. Their experiment results show that pose and edge cues help achieve better performance. However, the pose or edge in formations are not fully exploited. By exploring the correlation between the three factors, the HNL brings a boost of 2.23% mIoU to CE2P and 6.03% mIoU to MuLA. Even when removing the loss L_p, the 54.43% mIoU is higher than others, which indicates the direct supervision of the parsing encoder is necessary and our framework effectively utilizes pose and edge features to assist human body parsing. Moreover, the pose encoder in our network only consists of two deconvolution layers, and it is much simpler than the hourglass which is adopted in MuLA [29]. Thus, the performance may get higher if utilizing more powerful network.

ATR. Tab. 3 reports the results and comparisons with four recent approaches on ATR. The proposed method brings a significant performance gain in terms of every metric. Particularly, our model achieves 4.36% boost for F-1 score. This increase confirms the effectiveness of the pose and edge factors to parsing, and the correlation module has a strong capability to incorporate pose and edge information with the parsing features. Although the F-1 score 90.89% in [13] is higher than ours, it adopts DeepLabV3+ as backbone which is more complicated than ResNet101, and the input size 512×512 is larger than our 384×384. On the basis that the human joint labels are obtained from the output of the pose estimator [38], it illustrates that the proposed system is flexible and has a low-complexity to be deployed with no additional pose annotation cost.

Method	Backbone	mIoU
PGN [14]	ResNet101	55.80
Parsing R-CNN (R50) [39]	ResNet50	57.50
Graphonomy [13]	DeepLabV3+	58.58
Parsing R-CNN (X101) [39]	ResNeXt101	59.80
Ours	ResNet101	**60.18**

Table 4. Comparison of performance on the CIHP validation set.

4.2.3 Performance on multi-person datasets

CIHP. Experiment results are compared with the other approaches in Tab. 4 on the CIHP dataset. Our model outperforms the existing approaches and achieves 60.18 in terms of mIoU. The previous work [14] gets 55.80% mIoU by jointly conducting human parsing and edge detection. Parsing R-CNN [39] gets 57.50% mIoU using ResNet50 and its training images are in the size of 512×864. Using smaller input size and backbone ResNet101, our performance is 0.38% mIoU higher than Parsing R-CNN even when it changes the backbone to ResNeXt101. Our result is 1.6% mIoU higher than Graphonomy [13], which uses a graph convolution model and adopts a strong backbone DeepLabV3+ [6]. This performance suggests the superiority of our parsing method with the assistance of pose and edge factors, and correlating parsing with pose and edge can introduce contextual cues into human parsing task.

4.3. Evaluation of each component

We analyze the parameter sensitivities of our model in Tab. 6 and validate the effect of each component in Tab. 5.

The effect of different loss weights. The loss values in different branches are crucial to the model. In Tab. 6, we test four α values $\{0, 1, 2, 10\}$ with six β values $\{0, 1, 10, 50, 70, 80\}$. α = 0 or β = 0 indicates the baseline that removes the edge branch or pose branch from our model. It is observed that adding either edge or pose information to the parsing network brings a significant boost to the baseline. And the model achieves the highest mIoU when α = 2 and β= 70, which we choose as the final loss weights.

The effect of pose and edge cues. Firstly, we train a baseline model P which only contains parsing branch. In Tab. 5, without the contextual cues from pose and edge feature, the baseline model achieves 48.67% mIoU. We then add an edge/pose branch to the baseline model and concatenate parsing with edge/pose feature to perform prediction, denoted as P+B and P+K. Compared with baseline model

Method	hat	hair	glove	glass	u-clot	dress	coat	sock	pants	j-suit	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
DeepLabV2 [4]	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 [27]	57.66	65.63	30.07	20.02	64.15	28.39	51.98	41.46	71.03	23.61	9.65	23.20	69.54	55.30	58.13	51.90	52.17	38.58	39.05	84.75	46.81
SS-NAN [42]	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 [21]	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 [33]	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
P	63.61	69.18	36.25	27.68	67.23	31.80	53.69	43.45	71.75	28.76	14.33	24.39	72.33	57.76	60.74	47.80	47.38	34.18	34.90	86.22	48.67
PP	62.60	68.47	35.78	27.36	65.16	27.78	51.50	41.60	70.42	29.60	17.11	21.50	71.69	59.46	62.11	50.80	50.75	37.76	40.03	85.69	48.86
P+B	65.11	70.71	38.38	30.04	68.65	32.60	55.13	46.31	73.37	31.94	17.51	28.36	73.51	60.68	63.52	51.50	51.37	39.75	39.78	87.09	51.27
P+K	64.30	70.24	39.10	28.85	68.03	33.10	55.16	46.74	72.99	27.57	16.59	28.44	73.03	60.60	63.34	51.22	51.42	38.68	39.40	86.90	50.79
P+B+K	65.01	71.13	40.30	29.14	69.47	33.91	55.78	47.82	73.85	31.98	18.81	28.94	74.12	61.93	63.95	52.35	51.99	40.19	40.81	87.23	51.93
PB	65.43	71.77	40.69	26.00	69.32	32.82	56.33	46.61	74.52	30.87	23.46	27.51	74.28	64.23	66.68	57.64	56.72	44.80	44.80	87.77	53.11
PK	66.16	72.06	40.52	31.15	69.74	33.97	56.81	49.22	74.74	**32.56**	20.19	27.81	74.78	65.48	67.45	59.48	58.41	45.41	45.95	87.72	53.98
PBB	66.14	72.42	41.04	27.81	70.12	34.91	57.01	47.21	75.03	31.38	22.99	28.21	74.39	64.92	67.58	58.33	57.64	45.51	46.10	87.46	53.82
PKK	66.15	72.26	40.78	31.34	69.94	34.02	57.40	49.41	74.91	32.19	21.77	28.11	74.98	65.38	67.55	59.66	58.62	45.58	46.01	87.32	54.17
Ours (CorrPM)	**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**

Table 5. Comparison of per-class IoU on the LIP validation set. P: Only parsing feature; PP: Performing self-correlation on parsing feature; P+B/P+K: Concatenating parsing with edge/pose feature; P+B+K: Concatenating parsing, edge and pose feature; PB/PK: Correlating parsing with edge/pose feature; PBB/PKK: Correlating parsing with two edge/pose features. CorrPM outperforms existing methods and achieves 55.33% mIoU.

α	β					
	0	1	10	50	70	80
0	48.72	52.08	52.77	53.10	53.98	53.59
1	50.98	51.15	52.03	51.54	53.78	53.13
2	53.08	53.52	54.08	54.01	**55.33**	54.45
10	53.12	53.46	53.57	53.24	53.45	53.53

Table 6. Parameter discussion of α and β values in Equation 4 on the LIP dataset.

P, simple concatenation boosts 2.6% and 2.12% in terms of mIoU, respectively. Particularly, after fusing edge and parsing feature, the performances of some classes which are usually adjacent and have similar appearances (*e.g.*, upper clothes and pants), gain nearly 1.5% mIoU. These results demonstrate the effectiveness of edge and pose factors to parsing task. And the model P+B+K denotes concatenating both edge and pose feature with parsing feature. It only improves the performance by 0.66% mIoU compared with P+B, which indicates that even pose and edge factors are necessary for parsing, concatenating all three factors is not an ideal method to sufficiently leverage contextual cues.

The effect of self-correlation. To investigate the effect of the non-local operation, we add a traditional non-local self-attention module at the end of baseline model, denoted as PP. From Tab. 5, there is little improvement (0.09% mIoU) when calculating the relationship within parsing feature itself, and the performance of some classes is reduced such as hat, dress and upper-clothes. It shows that only exploiting the self-correlation of parsing feature is not enough and we need more pivotal factors from pose and edge to boost parsing performance.

The effect of correlation among parsing, edge and pose. We conduct two heterogeneous non-local correlations experiments, one is between parsing and edge factor, denoted as PB, and the other is between parsing and pose factor, denoted as PK, to validate the benefits of correlation module to parsing task. The performance improvement is more significant if leveraging the proposed heterogeneous non-local module, yielding 4.44% and 5.31% increases in terms of mIoU to baseline model P. And compared with P+B and P+K, the correlation module brings 1.84% and 3.19% mIoU gains. And even if we only use either pose or edge features, the result is more than 1.18% mIoU higher than the concatenation of all the three features, P+B+K. It is also observed that some categories which are closely related to human body joints are significantly improved by a large margin, which yields about 10% improvement in terms of mIoU. It shows that our HNL can make sufficient use of edge and pose information to accurately locate the boundary of semantics and maintain the body part geometry.

The effect of integration of multiple tasks. Aggregating the feature maps from multiple tasks will increase the channel number for fusion with parsing feature. Thus experiments are performed to study the efficacy of it. In Tab. 5, PBB (PKK) demonstrates the result of fusing two edge (pose) feature maps with parsing feature along channel dimension in HNL. PBB/PKK has the same channel number as CorrPM, while the mIoU is more than 1% lower than it. It shows the improvements are from the integration of multiple tasks rather than the increased channel number.

4.4. Qualitative Results

The solution of pose and edge to two problems. As mentioned in Sec. 1, there are two problems in human parsing task: inaccurate boundary localization between two adjacent parts and semantics inconsistency of segmented categories. Several images and sub-relation maps are shown in Fig. 4 to demonstrate the benefits that the proposed HNL learns from pose and edge information. The size of relation map S mentioned in Sec. 3.3 is $HW \times HW$. Hence, for a certain position in the image (marked as red point in Fig 4), the size of its corresponding sub-relation map is $H \times W$. As shown in the left half of Fig 4, some pixels in right arm are wrongly predicted as left arm, while there is no semantic boundary in this region. In the second row,

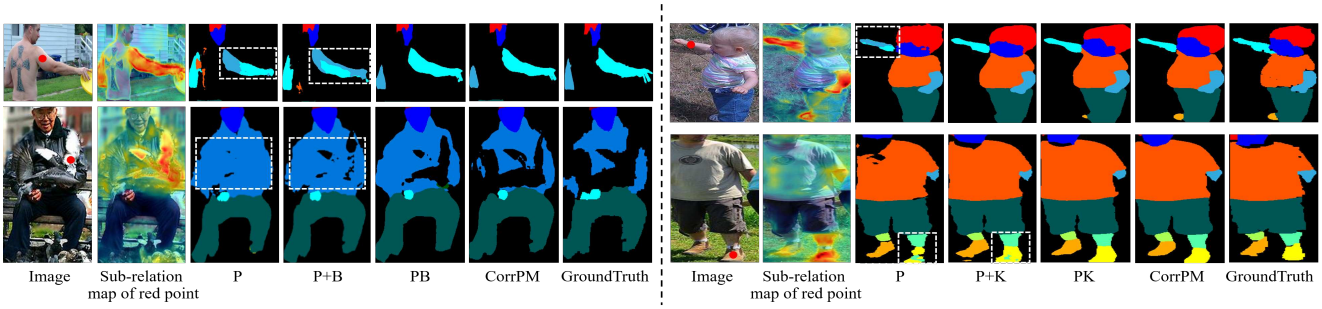

Figure 4. Visualization results of different fusion methods. These images show the benefits of edge/pose information to parsing task. The meaning of symbols is the same as Tab. 5.

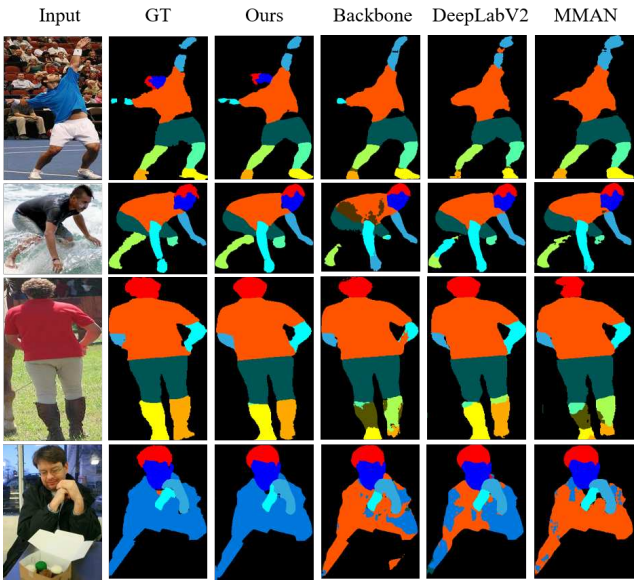

Figure 5. Visualization of different methods on the LIP dataset. The proposed CorrPM obtains smoother edge prediction and more reasonable body structure results.

the appearances of the coat and bird are similar so that the baseline model cannot tell them apart. After concatenating edge with parsing feature, the number of error pixels reduces but the boundary is still not clear. When utilizing correlation module, all the semantic edges are rightly predicted. Hence correlating edge with parsing factor can solve inaccurate boundary localization problem. From the right part of Fig. 4, the shoes region loses much detail during downsampling process, thus is not correctly classified. Concatenating pose with parsing feature can mitigate this problem. After correlating with parsing feature, the model obtains the awareness of the position of foot and shoe, hence the shoes classes are segmented correctly. Therefore, correlating pose with parsing factor can settle the semantics inconsistency matter.

Comparison with the previous methods. We show the quality results in Fig. 5 compared with DeepLabV2 [4], MMAN [27]. Our model outperforms other methods and the predictions are more precise. For example, on the first row, the head and right arms of the person are missing in other methods, while our model correctly predicts them despite the complexity of the background. Besides, with the help of edge information, our framework successfully locates the semantic boundary of the clothes and the legs shown in the second row, and keeps the semantics consistent among upper clothes category. We also observe from the third row that by adding pose information, the model can learn the global body structure of human and accurately identifies the left and right shoes, not legs. Consequently, the proposed HNL effectively employs the relationship of edge, pose and parsing features, and outputs more reasonable and precise results on the human parsing task.

5. Conclusion

In this paper, we propose a Correlation Parsing Machine (CorrPM) to take advantage of both semantic edge and human body keypoint features. For the two problems in human parsing task, our approach utilizes semantic edge to distinguish the boundary of two adjacent categories and human keypoint to enforce segmented classes to be consistent with body parts. With the heterogeneous non-local (HNL) module, the proposed model explores the relationship of edge, pose and parsing factors, and provides the contextual cues for human parsing task. The whole model is end-to-end learnable. Experiments on three benchmarks demonstrate the effectiveness of the proposed method. Moreover, the proposed system is flexible and easy to be deployed even without pose annotation.

Acknowledgments. This work is partially supported by the Beijing Major Science and Technology Project under contract No. Z191100010618003 and National Key Research and Development Program of China under contract No. 2016YFB0402001. We acknowledge Kingsoft Cloud for the helpful discussion and free GPU cloud computing resource support. We are also grateful to Dr Liang Zheng who is the recipient of an Australian Research Council Discovery Early Career Award (DE200101283) funded by the Australian Government.

References

[1] Vijay Badrinarayanan, Ankur Handa, and Roberto Cipolla. Segnet: A deep convolutional encoder-decoder architecture for robust semantic pixel-wise labelling. *Computer Science*, 2015.

[2] A. Buades, B. Coll, and J. M. Morel. A non-local algorithm for image denoising. In *IEEE Computer Society Conference on Computer Vision & Pattern Recognition*, 2005.

[3] Liang-Chieh Chen, Jonathan T Barron, George Papandreou, Kevin Murphy, and Alan L Yuille. 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*, pages 4545–4554, 2016.

[4] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE transactions on pattern analysis and machine intelligence*, 40(4):834–848, 2017.

[5] Liang-Chieh Chen, Yi Yang, Jiang Wang, Wei Xu, and Alan L Yuille. Attention to scale: Scale-aware semantic image segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3640–3649, 2016.

[6] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *Proceedings of the European conference on computer vision*, pages 801–818, 2018.

[7] Jifeng Dai, Kaiming He, and Jian Sun. Instance-aware semantic segmentation via multi-task network cascades. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3150–3158, 2016.

[8] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[9] Jian Dong, Qiang Chen, Wei Xia, Zhongyang Huang, and Shuicheng Yan. A deformable mixture parsing model with parselets. In *IEEE International Conference on Computer Vision*, 2014.

[10] Hao-Shu Fang, Guansong Lu, Xiaolin Fang, Jianwen Xie, Yu-Wing Tai, and Cewu Lu. Weakly and semi supervised human body part parsing via pose-guided knowledge transfer. *arXiv preprint arXiv:1805.04310*, 2018.

[11] Clement Farabet, Camille Couprie, Laurent Najman, and Yann LeCun. Learning hierarchical features for scene labeling. *IEEE transactions on pattern analysis and machine intelligence*, 35(8):1915–1929, 2012.

[12] Jun Fu, Jing Liu, Haijie Tian, Yong Li, Yongjun Bao, Zhiwei Fang, and Hanqing Lu. Dual attention network for scene segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3146–3154, 2018.

[13] Ke Gong, Yiming Gao, Xiaodan Liang, Xiaohui Shen, Meng Wang, and Liang Lin. Graphonomy: Universal human parsing via graph transfer learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7450–7459, 2019.

[14] Ke Gong, Xiaodan Liang, Yicheng Li, Yimin Chen, Ming Yang, and Liang Lin. Instance-level human parsing via part grouping network. In *Proceedings of the European Conference on Computer Vision*, pages 770–785, 2018.

[15] Ke Gong, Xiaodan Liang, Dongyu Zhang, Xiaohui Shen, and Liang Lin. 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*, pages 932–940, 2017.

[16] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017.

[17] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[18] Zilong Huang, Xinggang Wang, Lichao Huang, Chang Huang, Yunchao Wei, and Wenyu Liu. Ccnet: Criss-cross attention for semantic segmentation. In *The IEEE International Conference on Computer Vision*, October 2019.

[19] Dong Jian, Chen Qiang, Xiaohui Shen, Jianchao Yang, and Shuicheng Yan. Towards unified human parsing and pose estimation. In *2014 IEEE Conference on Computer Vision and Pattern Recognition*, 2014.

[20] Jianshu Li, Jian Zhao, Yunchao Wei, Congyan Lang, Yidong Li, Terence Sim, Shuicheng Yan, and Jiashi Feng. Multiple-human parsing in the wild. *arXiv preprint arXiv:1705.07206*, 2017.

[21] Xiaodan Liang, Gong Ke, Xiaohui Shen, and Lin Liang. Look into person: Joint body parsing and pose estimation network and a new benchmark. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PP(99):1–1, 2018.

[22] Xiaodan Liang, Chunyan Xu, Xiaohui Shen, Jianchao Yang, Si Liu, Jinhui Tang, Liang Lin, and Shuicheng Yan. Human parsing with contextualized convolutional neural network. In *Proceedings of the IEEE international conference on computer vision*, pages 1386–1394, 2015.

[23] Guosheng Lin, Chunhua Shen, Anton Van Den Hengel, and Ian Reid. Efficient piecewise training of deep structured models for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3194–3203, 2016.

[24] Tsungyi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollar, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European Conference on Computer Vision*, pages 740–755, 2014.

[25] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3431–3440, 2015.

[26] Xianghui Luo, Zhuo Su, Jiaming Guo, Gengwei Zhang, and Xiangjian He. Trusted guidance pyramid network for human parsing. In *2018 ACM Multimedia Conference on Multimedia Conference*, pages 654–662. ACM, 2018.

[27] Yawei Luo, Zhedong Zheng, Zheng Liang, Guan Tao, Junqing Yu, and Yang Yi. Macro-micro adversarial network for human parsing. In *European Conference on Computer Vision*, 2018.

[28] Alejandro Newell, Kaiyu Yang, and Jia Deng. Stacked hourglass networks for human pose estimation. In *European conference on computer vision*, pages 483–499. Springer, 2016.

[29] Xuecheng Nie, Jiashi Feng, and Shuicheng Yan. Mutual learning to adapt for joint human parsing and pose estimation. In *Proceedings of the European Conference on Computer Vision*, pages 502–517, 2018.

[30] Xuecheng Nie, Jiashi Feng, Yiming Zuo, and Shuicheng Yan. Human pose estimation with parsing induced learner. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018.

[31] Hyeonwoo Noh, Seunghoon Hong, and Bohyung Han. Learning deconvolution network for semantic segmentation. In *IEEE International Conference on Computer Vision*, 2015.

[32] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015.

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

[34] Ke Sun, Bin Xiao, Dong Liu, and Jingdong Wang. Deep high-resolution representation learning for human pose estimation. *arXiv preprint arXiv:1902.09212*, 2019.

[35] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7794–7803, 2018.

[36] Fangting Xia, Peng Wang, Liang-Chieh Chen, and Alan L Yuille. Zoom better to see clearer: Human and object parsing with hierarchical auto-zoom net. In *European Conference on Computer Vision*, pages 648–663. Springer, 2016.

[37] Fangting Xia, Peng Wang, Xianjie Chen, and Alan L Yuille. Joint multi-person pose estimation and semantic part segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6769–6778, 2017.

[38] Bin Xiao, Haiping Wu, and Yichen Wei. Simple baselines for human pose estimation and tracking. In *Proceedings of the European Conference on Computer Vision*, pages 466–481, 2018.

[39] Lu Yang, Qing Song, Zhihui Wang, and Ming Jiang. Parsing r-cnn for instance-level human analysis. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 364–373, 2019.

[40] Yulun Zhang, Kunpeng Li, Kai Li, Bineng Zhong, and Yun Fu. Residual non-local attention networks for image restoration. *arXiv preprint arXiv:1903.10082*, 2019.

[41] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2881–2890, 2017.

[42] Jian Zhao, Jianshu Li, Xuecheng Nie, Fang Zhao, Yunpeng Chen, Zhecan Wang, Jiashi Feng, and Shuicheng Yan. Self-supervised neural aggregation networks for human parsing. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 7–15, 2017.

[43] Zhen Zhu, Mengde Xu, Song Bai, Tengteng Huang, and Xiang Bai. Asymmetric non-local neural networks for semantic segmentation. In *The IEEE International Conference on Computer Vision*, October 2019.

[44] Yueqing Zhuang, Fan Yang, Li Tao, Cong Ma, Ziwei Zhang, Yuan Li, Huizhu Jia, Xiaodong Xie, and Wen Gao. Dense relation network: Learning consistent and context-aware representation for semantic image segmentation. In *International Conference on Image Processing*, pages 3698–3702, 2018.