This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2021.3108771, IEEE Transactions on Pattern Analysis and Machine Intelligence

On the Correlation among Edge, Pose and Parsing

Ziwei Zhang, Chi Su, Liang Zheng, Xiaodong Xie, Yuan Li

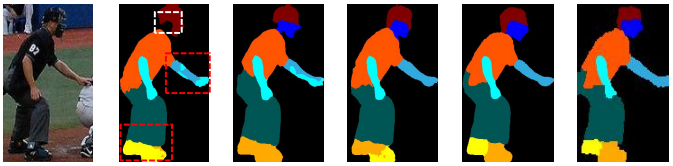
Abstract—Semantic parsing, edge detection and pose estimation of human are three closely-related tasks. They present human characteristics from three complementary aspects. Compared to learning them individually, solving these tasks jointly can explore the interaction of their contextual cues. However, prior works usually study the fusion of two of them, *e.g.*, parsing and pose, parsing and edge. In this paper, we explore how pixel-level semantics, human boundaries and joint locations can be effectively learned in a unified model. Specifically, we propose an end-to-end trainable Human Task Correlation Machine (HTCorrM) to implement the three tasks. It is asymmetric in that it supports a main task using the other two as auxiliary tasks. We also introduce a Heterogeneous Non-Local module (HNL) to discover the correlations of the three heterogeneous domains. HNL fully explores the global dependency among tasks between any two positions in the feature map. Experimental results on human parsing, pose estimation and body edge detection demonstrate that HTCorrM achieves competitive performance. We show that when designated as the main task, the accuracy of each of the three tasks is improved. Importantly, comparative studies confirm the advantages of our proposed feature correlation strategy over feature concatenation or post processing.

Index Terms—Correlation Machine, Heterogeneous Non-Local, Human Parsing, Pose Estimation, Human Body Edge Detection

1 Introduction

Analysis of human is a challenging problem in recent years [1], [2], [3], [4], [5], [6], and is widely used for applications in video forensics [7], human action recognition [8], [9], and human tracking [10], [11]. Among them, human parsing, pose estimation and human body edge detection are three fundamental and extensively-studied tasks. Although more and more attentions have been drawn to the three tasks, much previous researches are individually-learned methods which study one task without the help of other cues [2], [12], [13]. This may lead to mislabeling or missing predictions when context clues are not obvious. Take human parsing as an example. As depicted in Fig. 1(b), prior literature finds that there are two problems in individually-learned human parsing methods, *i.e.*, boundary confusion and human body structure inconsistency.

Recently, it is shown that the three tasks are essentially highly associated [14], [15]. As illustrated in Fig. 2, they present human characteristics from three complementary aspects. It is observed that human parsing explores the dense semantic contextual information. Pose estimation captures the relations of the connected keypoint and gives a rational body structure. Edge detection distinguishes the border between two parts and helps to remove the misleading predictions outside the human instance. When one task is individually learned, it may capture incomplete cue, thus encounters some problems as described above. But if we fuse multiple tasks, *e.g.,* correlate them, the model can obtain contextual cues from multiple domains, which serves as the supplementary features for one task. Motivated by this fact, researchers exploit one task to help other tasks to solve the existing problems [1], [5], [14], [15]. For instance, some state-of-the-art works fuse either edge or pose cue with parsing feature [1], [14], [15] by direct feature concatenation or post processing shown in Fig. 1(c) and (d). These fusion solutions can bring a certain degree of performance gain over the individually-learned methods.



(a) Image (b) Parsing baseline (c) Parsing and edge (d) Parsing and pose (e) Our method (f) GroundTruth

Fig. 1. Example of how parsing errors can be alleviated with assistance of pose and edge. (a) An input image. (b)-(e) are parsing results of various methods. (b) Parsing baseline [16]. (c) Fusion of parsing and edge through feature concatenation [1]. (d) Fusion of parsing and pose through post processing [15]. (e) Our method. (f) Parsing GroundTruth. (b) shows two types of parsing errors: boundary ambiguity (white box) and body structure inconsistency (red box). The fusion of boundary features (c) or keypoint features (d) may mitigate one of them. The two types of errors are both alleviated 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.

In spite of the progress made till date, contemporary methods have not leveraged the related cues to the full potential, thus have some drawbacks. First, they use *a single factor* to help another [1], [5], [17], which might be able to handle only *a single problem* mentioned above. For example, in Fig. 1(c), when using edge as an auxiliary task to parsing, the boundary between hat and face is relatively accurate, but the boundary between left shoe and right shoe

- *Corresponding author: Yuan Li.*
- *Z. Zhang, X. Xie and Y. Li are with the Department of EECS, Peking University, Beijing 100871, China. E-mail: {ziwei.zh, donxie, yuanli}@pku.edu.cn*
- *C. Su is with Kingsoft Cloud, Beijing 100085, China. Email: suchi@kingsoft.com*
- *L. Zheng is with the Research School of Computer Science, The Australian National University, Canberra, ACT 0200, Australia. E-mail: liangzheng06@gmail.com*

0162-8828 (c) 2021 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.
Authorized licensed use limited to: Peking University. Downloaded on January 12,2022 at 15:26:18 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2021.3108771, IEEE Transactions on Pattern Analysis and Machine Intelligence

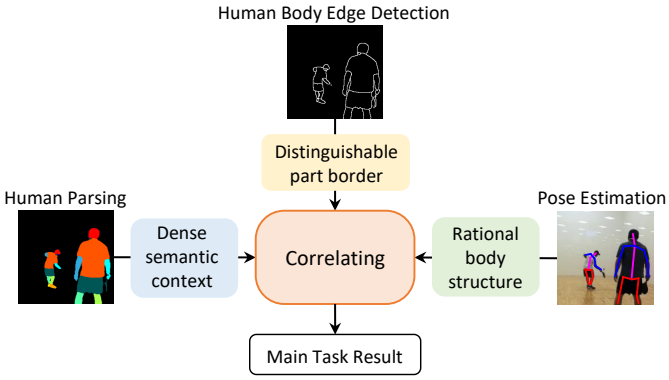


Fig. 2. Illustration of the complementary cues of three correlated tasks: parsing, edge and pose. They analyze human from three different perspectives and are highly associated. Human parsing aims to partition each pixel to a category, exploiting the semantic context. Human body edge detection extracts the distinguishable borders between adjacent parts to facilitate pixel-wise parsing and keypoint localization. Pose estimation predicts the locations of human body joints, providing rational body part structure. Correlating the three tasks can attain the complementary cues from the other two tasks to support the main task, which is specified from one of parsing, pose and edge.

is still erroneous. Second, the fusion strategies that existing research usually adopts is direct feature concatenation or post processing [15], [18], which neglects their long-range correlations on the feature map. Thus, such practice may be limited to generate optimal results.

To address above problems, we propose a Human Task Correlation Machine (HTCorrM) to simultaneously integrate the three tasks: parsing, pose and edge in a unified model. We implement each of them as the main task, which is supported by the other two auxiliary tasks. In this way, the proposed HTCorrM learns the human task with the support from three respects: pixel-level semantics, human boundaries and joint locations, and get the main task result, as shown in Fig. 2, which is more reasonable. Instead of using feature concatenation or post processing, HTCorrM explores the correlation among them to strengthen the model. It consists of a backbone, three feature encoders and a heterogeneous non-local (HNL) module. The backbone extracts the shared feature from multiple tasks. It can be either a specific architecture switched according to the main task, or a unified model that can be trained to suit three tasks. The feature encoders calculate the respective feature representations for the three interdependent tasks. HNL combines the features from three heterogeneous domains into a hybrid feature. It explores the spatial affinity between this hybrid feature and the feature map of the main task at all positions. As such, it can obtain the cues from the two auxiliary tasks and effectively aggregates them into the main task through correlation operation. In the example of Fig. 1(e), when parsing is switched to the main task, with the help of pose, HTCorrM perceives human body structure coherence, thus corrects the mislabeling of the shoes. In the same time, correlating parsing with edge can facilitate outlining the borders between the adjacent parts. So HTCorrM accurately locates the boundary between face and hat. Overall, it addresses boundary localization problem and human body structure inconsistency problem.

Moreover, we find that when performing multi-human analysis, capturing correlations on the mask level by pre-processing using Mask R-CNN [19] as in [20] exerts some negative effects (refer to Section 3.5 for more details). To overcome the drawbacks and capture task correlations on the image level instead of mask level, we design an end-to-end framework that yields significant performance improvement over the two-stage network [20]. Note that our method does not belong to Multi-Task Learning (MTL). We employ two auxiliary tasks to aid the main task by exploring their correlations, while MTL performs several tasks in parallel and neglects the synergy among them.

In summary, our contributions are three-fold. 1) We propose an end-to-end trainable Human Task Correlation Machine (HTCorrM) which integrates parsing, pose and edge in a unified model. 2) We design a Heterogeneous Non-Local (HNL) structure to explore the full-image correlation among the three tasks, one as the main, and the other two auxiliary. When being implemented as the main task, each one of them can benefit from the other two cues. 3) We report competitive accuracy on human parsing, edge detection and pose estimation tasks when any one of them is specified as the main task.

Compared with our conference version [20], we present the following new materials. 1) While [20] only supports human parsing as the main task, the proposed HTCorrM is generic that can effectively switch the main task to any of human parsing, pose estimation and edge detection. This further demonstrates the necessity of correlating one task with its related tasks. 2) While Mask R-CNN was used to extract human masks in [20], this paper improves the network to an end-to-end one for multi-human pixel-level analysis. This modification is non-trivial, as it allows us to capture among-task dependency from the entire image instead of single person masks, so that more reasonable and precise predictions are made. 3) More extensive experiments on public datasets LIP, ATR, CIHP and MPII are given to show the superiority of our proposed model.

2 Related Work

Human parsing. State-of-the-art methods in this area are based on deep learning. For example, Fully Convolutional Network [21], [22], [23] is an end-to-end solution and is widely used in later structures such as U-Net [24], SegNet [25], Encoder-Decoder network [26] and DeconvNet [27]. Aiming to enlarge the receptive field and obtain image contexts, DeepLab [16] and PSPNet [28] aggregate multi-scale object clues to make the segmentation more precise. The above are methods in general semantic segmentation. Focusing on segmenting human parts, Liang *et al.* [29] propose a Co-CNN framework capturing cross-layer local and global context information. Gong *et al.* [17] introduce a large-scale benchmark LIP and propose a self-supervised structure-sensitive learning method. In [30], Li *et al.* tackle human parsing problem by generating global parsing maps for human in a bottom-up way.

Pose estimation. Many works in human pose estimation leverage CNN to capture spatial context between keypoints [31], [32], [33], [34], [35]. Some representative CNN frameworks serve as bases for pose estimation task. Newell *et al.* [3] propose an Hourglass architecture, performing repetitive bottom-up, top-down procedures to explore multi-

0162-8828 (c) 2021 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.
Authorized licensed use limited to: Peking University. Downloaded on January 12,2022 at 15:26:18 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2021.3108771, IEEE Transactions on Pattern Analysis and Machine Intelligence

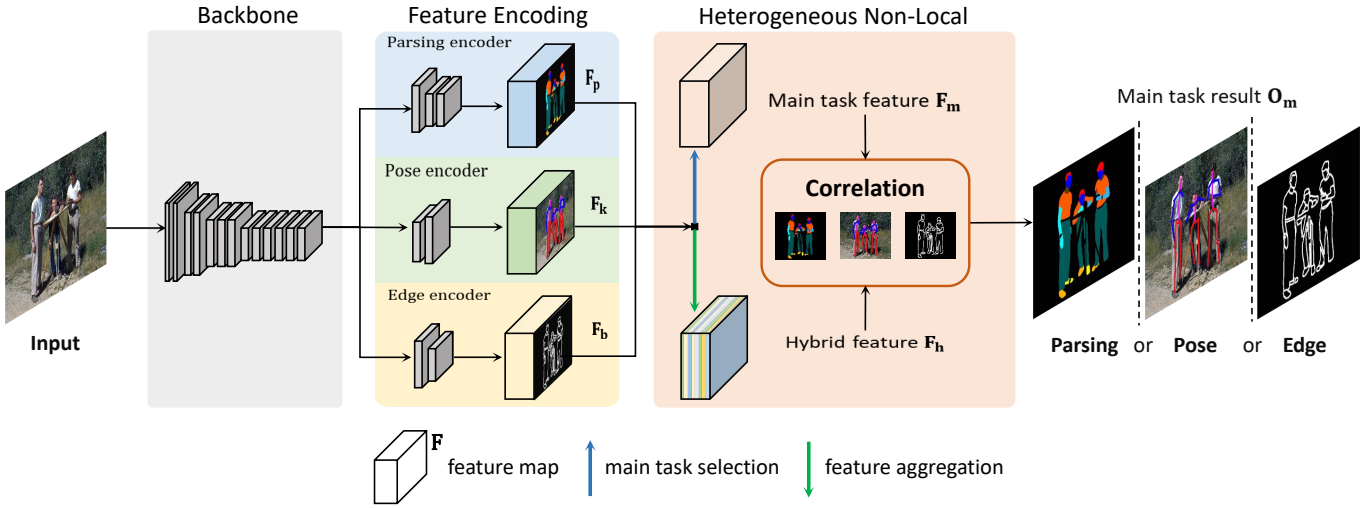


Fig. 3. Overall architecture of our system. It has a backbone, three feature encoders and a heterogeneous non-local (HNL) module. The backbone can be either task-specific or unified. It extracts the shared feature of the three tasks. The parsing, pose, edge encoders generate the corresponding features $\mathbf{F_p}$, $\mathbf{F_k}$ and $\mathbf{F_b}$. HNL aggregates these features into a hybrid representation $\mathbf{F_h}$, correlates it with the main task feature $\mathbf{F_m}$ (specified from one of the three features $\mathbf{F_p}$, $\mathbf{F_k}$ and $\mathbf{F_b}$), and obtains the main task result $\mathbf{O_m}$. For example, if human parsing is the main task, the main task feature $\mathbf{F_m}$ is the parsing feature $\mathbf{F_p}$, and it is correlated with $\mathbf{F_h}$. Finally, the network outputs the refined parsing map $\mathbf{O_m}$. The whole model is end-to-end trainable.

scale features. Xiao *et al.* [2] provide a simple baseline by adding several deconvolutional layers to the backbone, which achieves significant performance. To maintain high-resolution representations, Sun *et al.* [36] design a new architecture consisting of parallel multi-resolution subnetworks and fuse the outputs of them. Those architectures are widely adopted in pose estimation task because they provide a solid baseline to capture contextual information of body joint and achieve good performance on public datasets.

Edge detection. Traditional edge detectors [37], [38], [39] utilize low-level features to find maximum gradient magnitudes or rapid change in appearance as edges. These low-level cues need careful design and lack semantic information. Hence, many studies employ neural network to implicitly extract discriminative edge features recently. Xie *et al.* [40] learn representations in a deeply-supervised architecture and conduct end-to-end training. CASENet [41] adds skip-layer module and fuses features from different layers. He *et al.* [13] propose to enrich the multi-scale representations and introduce a bi-directional cascade model to let each CNN layer concentrate on a scale of edge.

Utilizing related factors to solve one task. Parsing, pose and edge provide complementary cues for each other [42], [6], [15], [17], [4], [43]. To help human parsing task, Chen *et al.* [44] propose an edge-aware filtering method to capture accurate semantic contours between two adjacent parts. Ruan *et al.* [1] fuse the edge map with parsing feature which can reserve the boundary of person parts to benefit human parsing. Gong *et al.* [45] conduct both semantic part parsing and edge detection in the way of sharing intermediate representation of both features. To improve pose estimation, Nie *et al.* [5] learn to adapt the parameters of pose model from parsing representations. Ladicky *et al.* [18] tackle joints occlusion problem with the help of pixel-wise body part information. When edge detection is the focus, Hu *et al.* [46] jointly perform edge detection and semantic segmentation tasks by a two-stream FCN. As aforementioned, these methods usually perform direct feature concatenation or post processing for feature refinement, which, as to be shown, is inferior to our correlation strategy of guiding the model to learn contextual cues. Moreover, we carefully design the branch of each task and their interaction modules, thus simultaneously integrating all the three tasks.

Attention. Our heterogeneous non-local block is an extension of standard non-local neural network [47], which utilizes the mechanism of attention. The attention model is leveraged in many high-level computer vision tasks to capture long-range dependencies. Wang *et al.* [47] propose the non-local block as a weighted summation of relationships of every position and show good performance in video classification. Vaswani *et al.* [48] propose a self-attention model for machine translation task. PSANet [49] learns the self-adaptive mask by the point-wise spatial attention operation. We note that previous methods mainly seek the relationship *within only one type of feature* in a homogeneous manner [50], [51], [52], [53]. In comparison, our proposed HNL explores a heterogeneous usage of the non-local network. It aggregates features from parsing, edge and pose, acquiring reliable feature representations to aid the main task. We later show that HNL can boost each of the three tasks after switching it to the main task and achieves competitive performance with the help of the other two auxiliaries.

3 Proposed Approach

We propose a unified Human Task Correlation Machine (HTCorrM) to solve human parsing, pose estimation and body edge detection tasks by specifying one as the main task and the other two as auxiliary. Given an input image $I \in \mathbb{R}^{3\times M\times N}$, our goal is to predict the parsing or pose or edge result depending on the main task. We have three types of labels: human body part category $\mathcal{P} \in \{0, 1, ..., Q-1\}^{M\times N}$, semantic boundary $\mathcal{B} \in \{0, 1\}^{M\times N}$ and human body keypoint location $\mathcal{K} = \{(x_i, y_i)\}_{i=1}^{J}$. Q and J are the number of part categories and body joints. (x_i, y_i) are the coordinates

0162-8828 (c) 2021 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.
Authorized licensed use limited to: Peking University. Downloaded on January 12,2022 at 15:26:18 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2021.3108771, IEEE Transactions on Pattern Analysis and Machine Intelligence

of keypoint i. The pixel belongs the boundary is labeled as 1, and the rest of them are labeled as 0.

The architecture of HTCorrM is shown in Fig. 3. It consists of a backbone, three feature encoders and is featured by a heterogeneous non-local (HNL) module. We obtain the result $\mathbf{O_m} \in \mathbb{R}^{T \times M^{'} \times N^{'}}, T \in \{Q, J, 2\}$ of the main task. In the following sections, we first demonstrate each component of HTCorrM in Section 3.1 (the backbone that can be either task-specific or unified), Section 3.2 (three feature encoders) and Section 3.3 (heterogeneous non-local module). Then the training objective is introduced in Section 3.4. Finally, Section 3.5 presents some discussions to help get a further understand of HNL.

3.1 Backbone

A backbone extracts features shared among the three tasks. It is either specific to the main task or unified. On the one hand, human parsing, pose estimation and edge detection analyze human from different perspectives, and thus their underlying problems are different. Therefore, given a main task, a main-task specific backbone is used. We show that our system can be conveniently switched to such specific backbones. In more details, we use DeepLabV2 [54], HRNet [55] and BDCN [13] as the backbones for parsing, pose and edge tasks, respectively. DeepLabV2 provides large receptive fields for parsing, HRNet designs high-resolution representations for pose, and BDCN precisely captures multi-scale body part edges. On the other hand, the three tasks we study are highly correlated, so we explore the possibility of a unified backbone applied to whichever is specified as the main task. In our work, we choose HRNetV2-W48 [55] as the unified backbone, as it is shown to be able to maintain a strong position sensitivity and learn high-resolution representations for dense prediction tasks [55].

3.2 Feature Encoders

Parsing encoder. Image context is captured in many previous works [16], [28] and is also essential in human parsing. We use Atrous Spatial Pyramid Pooling (ASPP) [16] to obtain more useful context cues and enlarge the receptive fields. It consists of five parallel branches: a global average pooling layer, a 1×1 convolution layer and three 3×3 atrous convolution layers with rates of (12, 24, 36). After that, all feature maps are concatenated into one feature whose channel number is reduced by 1×1 convolution layer. Meanwhile, some objects (e.g., sunglasses and socks) are relatively small and have a low resolution in the image, so their details might be lost during downsampling. We upsample the output of ASPP to the same resolution as the feature of lower stage G_1 in the backbone (*e.g.*, *Res2* in DeepLabV2) and concatenate them, which gets the parsing feature $\mathbf{F_p}$. This parsing encoder outputs a coarse semantic map and is supervised by pixel-wise prediction error.

Pose encoder. We adopt the idea from [2] to implement a simple but effective pose encoder only consisting of two transposed convolution layers. The kernel sizes of these two layers are both set to 4 with a stride of 2. Consequently, the shared feature of the backbone is upsampled, and we obtain the pose feature $\mathbf{F_k}$ with the same scale as $\mathbf{F_p}$.

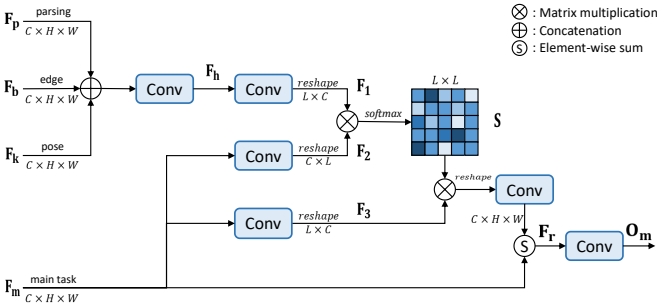


Fig. 4. The Heterogeneous Non-Local (HNL) module. It aggregates parsing, edge and pose features $\mathbf{F_p}$, $\mathbf{F_b}$ and $\mathbf{F_k}$ into a hybrid feature $\mathbf{F_h}$. Then it calculates the relationship between $\mathbf{F_h}$ and the main task feature $\mathbf{F_m}(\mathbf{m} \in \{\mathbf{p}, \mathbf{k}, \mathbf{b}\})$, producing the relation map $\mathbf{S}$. Symbols $\mathbf{F_1}, \mathbf{F_2}$ and $\mathbf{F_3}$ represent intermediate feature maps, and we utilize $\mathbf{F_r}$ as the refined feature to generate the final main task result $\mathbf{O_m}$.

After $\mathbf{F_k}$ is obtained, we regress the pose heatmap from it. Following [36], the groundtruth heatmap is generated by applying a 2D Gaussian filter centered on each annotated keypoint coordinate with standard deviation of 2 pixels. L2 loss is used as supervision. Here, we note that while Hourglass [3] is also a widely used structure [3], [14], [15], we do not consider it because of its higher complexity and computational cost.

Edge encoder. Bottom layers in the backbone are sensitive to local patterns, and top-layer feature obtains semantic consistency [24]. To extract accurate body part edge representations with semantics, features in both lower and higher stages are employed. We take the features of G_1, G_2, G_3 from backbone (more details in Section 4.2) into three 1×1 convolution layers to adjust their feature channels to be the same. All the feature maps are upsampled to the same size by linear interpolation. Then, they are concatenated and fed into one 3×3 and one 1×1 convolution layer successively to generate the edge feature map $\mathbf{F_b}$. The edge encoder is trained under the supervision of the difference between the predicted edge map and the groundtruth.

After features $\mathbf{F_p}, \mathbf{F_k}$ and $\mathbf{F_b}$ have been extracted, they will be fed into the heterogeneous non-local module to further explore the correlated guidance from three cues to boost the main task's performance.

3.3 Heterogeneous Non-Local Module

The associated tasks can provide complementary cues to each other. Recently, the attention module such as the non-local block is widely used to capture the long-range contextual information [50], [51], [52]. However, existing methods follow a mechanism of self-attention [47], which only accepts one input from one domain. Therefore, it is homogeneous and fails to exploit the correlations among multiple tasks. In contrast, we propose a Heterogeneous Non-Local (HNL) module that simultaneously aggregates the related cues from three heterogeneous domains, and it captures the global dependency among tasks between any two positions in the feature map.

Fig. 4 illustrates the detailed structure of HNL. $\mathbf{F_p}$, $\mathbf{F_k}$ and $\mathbf{F_b}$ are parsing, pose and edge features with the same dimension $C \times H \times W$. We first aggregate all the three features by concatenating them along the channel dimension,

0162-8828 (c) 2021 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.
Authorized licensed use limited to: Peking University. Downloaded on January 12,2022 at 15:26:18 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2021.3108771, IEEE Transactions on Pattern Analysis and Machine Intelligence

then a convolution layer parameterized by $\mathbf{W_1}$ is used to transform it into a hybrid feature $\mathbf{F_h} \in \mathbb{R}^{C\times H\times W}$:

$$\begin{aligned}\mathbf{F} &= concat(\mathbf{F_p}, \mathbf{F_k}, \mathbf{F_b}),\\ \mathbf{F_h} &= \mathbf{W_1}\mathbf{F}.\end{aligned} \tag{1}$$

We then specify one of parsing, pose and edge as the main task and represent its feature as $\mathbf{F_m}(\mathbf{m} \in \{\mathbf{p}, \mathbf{k}, \mathbf{b}\})$. Instead of using the self-attention mechanism as the standard non-local block does, we perform heterogeneous correlation operation between the hybrid feature $\mathbf{F_h}$ and main task feature $\mathbf{F_m}$. First, $\mathbf{F_h}$ and $\mathbf{F_m}$ are fed into two convolution layers to generate two intermediate features $\mathbf{F_1}$ and $\mathbf{F_2}$. We reshape and transpose them into matrixes of size $L \times C$ and $C \times L$, respectively, where $L = H \times W$ denotes the total number of pixels per channel.

$$\mathbf{S} = softmax(\mathbf{F_1} \cdot \mathbf{F_2}), \tag{2}$$

where a point (i, j) in $\mathbf{S}$ measures the relation affinity between the i^{th} pixel in hybrid feature $\mathbf{F_h}$ and the j^{th} pixel in main task feature $\mathbf{F_m}$.

$\mathbf{F_m}$ is later delivered into another convolution layer, reshaped and transposed to $\mathbf{F_3} \in \mathbb{R}^{L\times C}$. We multiply it by $\mathbf{S}$ to transmit the correlation cues from all the associated tasks to the main task. The resulting feature is sent to a convolution layer parameterized by $\mathbf{W_2}$ and added back to $\mathbf{F_m}$ element-wisely to get the feature $\mathbf{F_r}$. $\mathbf{F_r}$ is leveraged as a refined feature and fed into a 1×1 convolution layer $\mathbf{W_3}$ to get the final result of the main task $\mathbf{O_m}$. The overall procedure can be formulated as:

$$\begin{aligned}\mathbf{F_r} &= \mathbf{W_2}(\mathbf{S} \cdot \mathbf{F_3}) + \mathbf{F_m},\\ \mathbf{O_m} &= \mathbf{W_3}\mathbf{F_r},\end{aligned} \tag{3}$$

where $\mathbf{W_2}$ is initialized as 0. Eq. 3 is a process of feature refinement, utilizing the relationship among the three tasks. In this way, HNL effectively aggregates the associated cues together. For instance, if we conduct HNL on human parsing, whose complementary tasks are pose and edge factors, $\mathbf{F_r}$ will obtain edge attention between two bordered parts and retain semantic consistency with human body, thus can get more reasonable parsing result $\mathbf{O_m}$.

HNL is an extension of non-local neural network [47] and has several advantages. First, if we follow the practice of the standard non-local network that calculates the pair-wise relationships of n tasks, it may lead to high computation complexity $O(n^2)$ and hard-convergence problem. However, the proposed HNL aggregates these three features into a hybrid feature and calculates the relationship between it and the main task feature, which is more efficient. Second, it integrates three complementary tasks into a unified model, while existing studies only fuse the two of them, so they cannot fully leverage the relations and can barely solve *one single problem* mentioned in Section 1. Finally, previous fusion methods [1], [5] also have refinement process such as feature concatenation. But they are simple and the refined feature can not capture the among-task dependency. Compared with it, HNL proposes a more effective correlation strategy which substantially delves the powerful cues so as to boost the model.

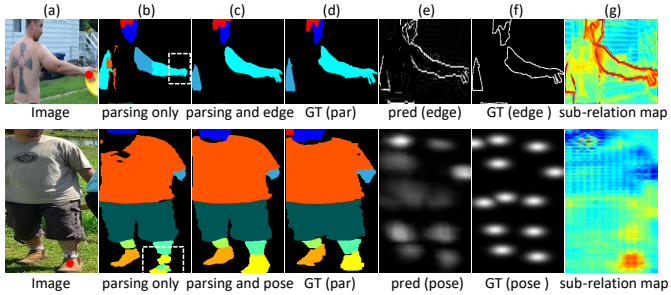


Fig. 5. Examples visualizing how auxiliary tasks benefit the main task. Two example models are used: correlating parsing and edge (top row), and correlating parsing and pose (bottom row). We specify two query points in Column (a): a point on the border of *arm* (top row), and a point on *left shoe* (bottom row), and show the corresponding relation maps produced by HNL in Column (g). Column (d) is the groundtruth of parsing and Column (e) is the prediction of edge (or pose). The parsing only baseline in (b) has erroneous predictions highlighted in white boxes. As auxiliary tasks, edge extracts borders between adjacent parts, and pose provides rational body part structure to the main task. By correlating parsing feature with edge / pose feature, the higher response areas (marked in red) in the relation map are discovered that contribute to the feature of the query (red point). Through the weighted sum in Eq. 3, the proposed model aggregates main-task features with the help of the complementary auxiliary tasks and eliminates errors of the baseline.

3.4 Training Objective

Three types of annotations from the related tasks are demanded to train the whole model, including the parsing supervision $\mathcal{P}$, human keypoint location $\mathcal{K}$ and semantic part boundary $\mathcal{B}$. The total training objective is:

$$\mathbf{L} = \mathbf{L_m} + \alpha\mathbf{L_p} + \beta\mathbf{L_k} + \gamma\mathbf{L_b}, \tag{4}$$

where α, β and γ are the loss weights to balance different tasks. $\mathbf{L_m}$ is the loss between the result $\mathbf{O_m}$ and the label of the main task, and $\mathbf{L_p}, \mathbf{L_k}$ and $\mathbf{L_b}$ are the losses between three encoders' outputs and their corresponding groundtruth. We apply different loss functions to different tasks. Cross Entropy loss is adopted to parsing task and edge detection task, and Mean Square Error loss is used for pose estimation task. The loss of one task keeps the same when it is specified as the main task or auxiliary. The whole framework is trained end-to-end.

3.5 Discussions

We discuss how HNL utilizes the auxiliary tasks to benefit the main task, and compare our method with: 1) self-attention network, 2) multi-task learning, and 3) some feature fusion strategies. Some drawbacks of Mask R-CNN pre-processing in the previous conference paper [20] are also summarized when performing multi-human parsing. We compare the visualized results in Fig. 6 to explain the importance of computing correlations on the image level instead of mask level. We also discuss the required manual labels in our method.

How does HNL utilize the auxiliary tasks to benefit the main task? To answer this question, we show two examples in Fig. 5, where we use human parsing as main task and deploy two models: 1) correlating parsing and edge, and 2) correlating parsing and pose. We visualize their relation maps which are compared with the baseline model (parsing only). Note that the size of the relation map is $HW \times HW$;

0162-8828 (c) 2021 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.
Authorized licensed use limited to: Peking University. Downloaded on January 12,2022 at 15:26:18 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2021.3108771, IEEE Transactions on Pattern Analysis and Machine Intelligence

TABLE 1
Comparison of feature fusion methods on human parsing. "EA" represents solving the edge ambiguity problem and "BI" represents solving the boundary inconsistency problem. Prior 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 superior accuracy to other fusion strategies. Note that "Accuracy" is for indication purpose only; please refer to Section 4.3 for specific numbers.

Method	EA	BI	Fusion Strategy	Accuracy
[1]	✓		Feature concatenation	++
[45]	✓		Feature concatenation	++
[14]		✓	Parameters mutual learning	++
[17]		✓	Loss constraint	+
[15]		✓	Post processing	+
Ours	✓	✓	Correlation	+++

for visualization convenience, we select one HW map corresponding to one query point, and reshape it into $H \times W$. We mark the query points in red in Column (a). As shown in the first row of Fig. 5(b), the baseline fails to distinguish the border of *arm*. When the auxiliary task edge is added and correlated with parsing, the reasonable edge predictions in Fig. 5(e) indicate the learned edge feature is relatively discriminative. Therefore, correlating the edge feature with the feature on the red point will lead to higher response on the edges (marked in red) in Fig. 5(e), which advises the edge location to the parsing feature. So the output around *arm* in Fig. 5(c) is more accurate than the baseline. Overall, the proposed HNL captures the cross-modality relation to benefit the main task. The auxiliary cues serve as a guide to the main task: each pixel in the main task feature map will be able to aggregate pixels with closer semantics guided by the auxiliary feature. By utilizing the contextual cues provided by the auxiliary task, the main task gets benefits from it and the proposed model outputs more reasonable prediction.

Comparison with self-attention network. Our correlation strategy and self-attention network [47] both compute pixel-wise affinity. However, self-attention network calculates the affinity of one type of feature. Different from it, HNL calculates the correlation among three different tasks. Moreover, HNL does not add much computation complexity compared with self-attention structure while maintains competitive accuracy.

Comparison with multi-task learning. HTCorrM and multi-task learning have something in common. Both of them perform multiple tasks through several feature encoders. But there are two differences. 1) Multi-task learning treats each task equally by a few parallel networks, while HTCorrM focuses on one main task with the help of correlated features. 2) Multi-task learning does not take the correlation into consideration and the synergy is not exploited. In contrast, HTCorrM uncovers the correlation among multiple tasks and leverages the pivotal contextual cues to benefit the main task.

Comparison with some feature fusion strategies. Some works that aggregate pose or edge information to assist human parsing are compared with our methods in Table 1. For the edge ambiguity issue mentioned in introduction, CE2P [1] and PGN [45] fuse edge feature with parsing feature by concatenation. But this fusion strategy is simple and the useful cue from edge is not fully exploited. Aiming to solve the body inconsistency problem, MuLA [14] conducts two parallel human pose estimation and human parsing networks and mutually learns the parameters. But the training process is somewhat complicated. Meanwhile, Xia *et al.* [15] adopt FCRF as a way of post processing and Gong *et al.* [17] introduce a body joint loss to assist human parsing. All the above fusion methods employ a single factor and thus can merely handle a single problem. In comparison, our model combines parsing with both pose and edge. The experimental results also show a better accuracy compared with other listed methods, suggesting that exploring the correlation among the three factors is superior to previous feature fusion strategy.

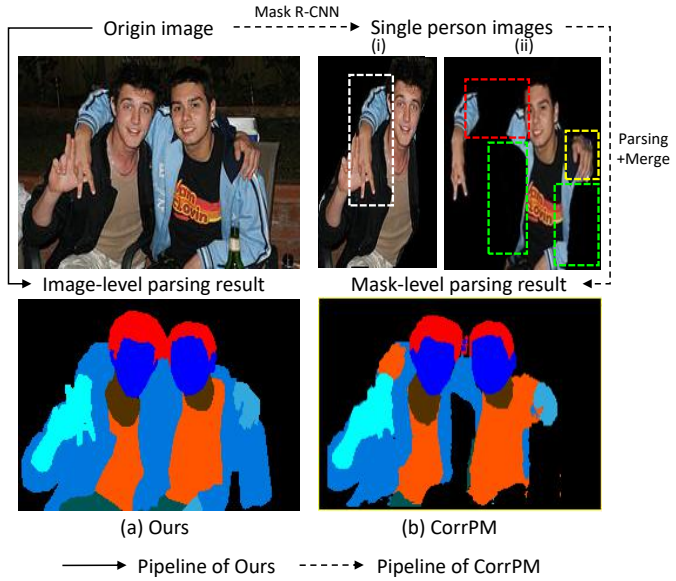


Fig. 6. Qualitative comparisons between CorrPM (our conference version) [20] and this paper. Our method extends [20] to an end-to-end model. Specifically, in this paper, we input the entire image into the network and output human parsing map directly. In comparison, the conference version [20] uses Mask R-CNN to generate single human images (i) and (ii), performs single human parsing and then merges the parsing results. In fact, when people are close (*e.g.*, hug each other), unsatisfying results may be yielded: 1) one person's limb can be wrongly segmented as someone else's (yellow box); 2) some parts maybe lost (green box); 3) during merging, duplicate predictions in one pixel will cause pixel ambiguity (white box). Instead of capturing the mask-level relationships, we remove Mask R-CNN and explore image-level correlations in an end-to-end fashion which avoids the above problems. So the network can acquire task dependency from the entire image, and the prediction cannot be harmed by the occluded body parts (missed by Mask R-CNN in red box).

Limitations of using Mask R-CNN to produce single person images. In the conference version [20], we first utilize Mask R-CNN [19] to generate the images of all single persons, perform single human parsing, and merge the parsing results of all the persons to produce the final parsing map. However, three limitations of such practice exist. First, the parsing results are highly dependent on the single person images. As shown in Fig. 6, after produced by Mask R-CNN, left arm of the left person is missing so this part is inaccurately classified as the background. Second, when merging the parsing results, there may have multiple predictions in one pixel which causes ambiguity. Third, the whole system is not end-to-end trainable. For this reason, we remove Mask R-CNN, conducting correlating operation on the whole image and the performance outperforms the con-

0162-8828 (c) 2021 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.
Authorized licensed use limited to: Peking University. Downloaded on January 12,2022 at 15:26:18 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2021.3108771, IEEE Transactions on Pattern Analysis and Machine Intelligence

ference paper's. It is because the end-to-end model avoids above separate-and-merge process, therefore, above three problems are overcome. On the other hand, some persons may be occluded by others in the multi-person images. As shown in the red box in Fig. 6, the pixels belonging to occluded parts are assigned as zeros. So the network cannot learn valuable information. Whereas, if we feed the whole image into the network, it can acquire the relations of all the people and capture task dependency from the entire image instead of the single person masks, and generates more reasonable prediction.

On manually annotated labels required in our method. We emphasize that not all the groundtruths used in Eq. 4 have to be manually labeled. In fact, as mentioned in Section 4.2, none of the datasets that we use have manual edge labels, and we generate such labels directly from the manual parsing labels by identifying the part boundaries. Moreover, some datasets only have manual annotations for parsing (pose), so we adopt pre-trained pose estimator [2], [56] (human parser [20]) to generate pose (parsing) labels. Therefore, the manually annotated samples only account for one third or at most two thirds of the entire dataset. It indicates our requirement of manual labels is not significantly heavy, the extent to which resembles semi-supervised algorithms [57]. For example, Oliver *et al.* use 8,000 out of 50,000 manually labeled data for image classification task. It would be interesting to see how semi-supervised algorithms can work under the context of human analysis, but because this paper focuses on effectively correlating multiple tasks, implementing semi-supervised add-ons seems out of the current scope.

4 Experiments

We evaluate our method on the LIP dataset [17], the ATR dataset [29], the CIHP dataset [45], and the MPII dataset [58]. The results show that our model achieves competitive performance on human parsing, pose estimation and edge detection tasks when they are specified as the main task, respectively. In the following sections, we present the datasets and detailed experimental settings in Section 4.1 and Section 4.2. Section 4.3 compares our model with some competing methods on the three tasks. Finally, we analyze the effect of components in HTCorrM in Section 4.4.

4.1 Datasets and Metrics

LIP is a single-human parsing benchmark focusing on human body parts and clothes labels. It contains both human pose and parsing groundtruth, including 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. We evaluate the performance of human parsing, pose estimation and edge detection tasks on it.

ATR 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 [29], we conduct experiments on it for single-human parsing utilizing 16,000 images for training, 1,000 for testing and 700 for validation.

(a) Image (b) Parsing label (c) Edge label

Fig. 7. MPII does not provide parsing and edge groundtruths, so we generate them to train HTCorrM that implements pose estimation as main task. Examples are shown in this figure. (a) Image. (b) Parsing labels generated by [20]. (c) Edge labels as the boundaries between two adjacent part categories in (b). We observe that they are reasonable and relatively accurate to provide reliable supervision.

CIHP is a multi-human parsing dataset providing 38,280 images with 19 semantic part labels and one background category. There are more than one person in each image. The dataset is collected from real-world scenes, including various viewpoints, resolutions and persons of challenging poses. It contains 28,280 training, 5,000 validation and 5,000 test images.

MPII is a large-scale benchmark tasked for human pose estimation. It provides 40k human samples which are split into 28k images for training and 11k for validation. There are 16 body joints annotated: *left/right ankle, left/right knee, left/right hip, pelvis, thorax, upper neck, head top, left/right wrist, left/right elbow*, and *left/right shoulder*.

We use Pixel Accuracy, Mean Accuracy, mIoU, Accuracy, Precision, Recall and F-1 score to evaluate parsing performance. For pose estimation, PCKh is adopted as the evaluation metric following [3]. For edge detection, we use F-measure at Optimal Dataset Scale (ODS) and Optimal Image Scale (OIS) following [13].

4.2 Implementation

Dataset annotation. According to Eq. 4, our experiments need labels with three types of annotations: pixel-wise parsing label, human body joint coordinates and edges between two semantic parts, but not all these annotations are required to be manually labeled. ATR and CIHP datasets only contain manual annotations for parsing, so we utilize pose estimator [2] and [56] pre-trained on COCO [59] to generate pose labels. Additionally, MPII only has the label of body joints, so the parsing annotations are generated by the algorithm [20]. Moreover, none of the four datasets have manual edge groundtruths, which are obtained from the parsing labels by calculating the border between two adjacent body part categories. We give two examples of parsing and edge labels in Fig. 7. The generated labels do not need additional annotation cost, and they are relatively accurate to bring performance improvement.

Training details. When human parsing is the main task, we apply two backbones: DeepLabV2 [16] with the

0162-8828 (c) 2021 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.
Authorized licensed use limited to: Peking University. Downloaded on January 12,2022 at 15:26:18 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2021.3108771, IEEE Transactions on Pattern Analysis and Machine Intelligence

TABLE 2
Method comparison on the single-human parsing dataset LIP. Per-class IoU (%) and mean IoU (%) are shown. "Ours (DeepLabV2)" and "Ours (HRNetV2-W48)" denote our method utilizing two backbones DeepLabV2 and HRNetV2-W48, respectively. Both are seen to be effective and achieve competitive accuracy. Using HRNetV2-W48 as backbone outperforms the DeepLabV2 backbone on 14 out of 20 classes, *e.g.*, *glove*, *glass* and *scarf*. "*" denotes our implementation. Comparing with [55], we use a smaller resolution of training images (384×384 *vs.* 473×473), and we do not use the augmentation strategy (*e.g.*, flip testing). So our implementation is slightly lower than reported.

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 [22]	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 [16]	56.48	65.33	29.98	19.67	62.44	30.33	51.03	40.51	69.00	22.38	11.29	20.56	70.11	49.25	52.88	42.37	35.78	33.81	32.89	84.53	44.03
Attention+SSL [17]	59.75	67.25	28.95	21.57	65.30	29.49	51.92	38.52	68.02	24.48	14.92	24.32	71.01	52.64	55.79	40.23	38.80	28.08	29.03	84.56	44.74
MMAN [12]	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 [23]	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
MuLA [14]	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	49.30
JPPNet [60]	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 [1]	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
HRNetV2*	65.70	71.21	40.86	29.27	68.24	36.66	55.50	45.93	73.63	24.75	23.66	25.54	74.42	64.85	68.04	58.92	59.07	47.85	48.55	86.87	53.48
CorrPM [20]	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
HRNetV2 [55]	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	55.90
Ours (DeepLabV2)	66.94	72.30	41.81	31.90	**70.95**	**38.47**	**58.80**	49.14	75.92	**33.15**	24.53	**29.99**	75.12	67.28	69.33	63.54	63.56	49.77	50.54	**88.53**	56.08
Ours (HRNetV2-W48)	**67.91**	**73.27**	**45.23**	**34.91**	70.76	37.49	57.43	**51.29**	**76.64**	30.25	**27.53**	28.80	**75.97**	**68.44**	**70.14**	**64.80**	**64.29**	**51.28**	**52.03**	**88.53**	**56.85**

ResNet101 [54] architecture, and HRNetV2-W48 [55]. Both are pre-trained on ImageNet [61]. For DeepLabV2, *Res2*, *Res3*, *Res4* are used as G_1, G_2 and G_3, respectively, mentioned in Section 3.2. For the unified backbone HRNetV2-W48, in all the experiments, its stage 1, stage 3, and stage 4 are adopted as G_1, G_2 and G_3, respectively. We train the network from scratch for 180 epochs. During training, the 384×384 input images are randomly rotated (from -60° to 60°), flipped and scaled (from 0.75 to 1.25). $\mathbf{F_p}, \mathbf{F_k}, \mathbf{F_b}$ 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 to 1e-3. Following previous works [62], 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 loss weights are set as follows: $\alpha = 1, \beta = 50, \gamma = 2$.

For pose estimation, we deploy HRNetV2-W32 and HRNetV2-W48 as the backbone. We fix the input size as 256×256. For data augmentation, we follow the same setting as [36], including random rotation from -45° to 45°, random scaling from 0.65 to 1.35 and random flipping. The network is trained for 220 epochs. We set the base learning rate to 1e-3 and it is decreased by 10 times at epoch 100, 150, 170 and 210. The feature size $C \times H \times W$ in HRNetV2-W32 and HRNetV2-W48 is $32 \times 64 \times 64$ and $48 \times 64 \times 64$, respectively. The loss weights are set as follows: $\alpha = 0.001, \beta = 1, \gamma = 0.0025$.

For edge detection, two backbones BDCN [13] and HRNetV2-W48 are deployed. For the task-specific backbone BDCN, we design ID block2, ID block3 and ID block4 as G_1, G_2 and G_3. SGD is used as optimizer and the whole network is trained for 40k iterations. The initial learning rate is set to 1e-6 and we decrease it by 10 times after every 10k iterations. The process of data augmentation is the same as [13]. The input size is fixed as 384×384. The loss weights are set as follows: $\alpha = 0.08, \beta = 200, \gamma = 1$.

Testing details. During testing, the output of HNL is adopted as the final result of the main task, while the output of the feature encoders are ignored. The inference procedure is conducted on a single scale by keeping the original aspect ratio and setting the long side of the image to 384 pixels (parsing and edge tasks) and 256 pixels (pose task). Multi-scale and flipping augmentation are not used.

TABLE 3
Method comparison on single-human parsing dataset ATR in terms of Accuracy (Acc), Foreground Accuracy (F.g.Acc), Precision (Pre), Recall (Rec) and F-1 score.

Method	Acc	F.g.Acc	Pre	Rec	F-1 score
Yamaguchi [63]	84.38	55.59	37.54	51.05	41.80
Paperdoll [64]	88.96	62.18	52.75	49.43	44.76
ATR [65]	91.11	71.04	71.69	60.25	64.38
DeepLabV2 [16]	94.42	82.93	69.24	78.48	73.53
Attention [22]	95.41	85.71	81.30	73.55	77.23
CoCNN [29]	96.02	83.57	84.59	77.66	80.14
TGPNet [66]	96.45	87.91	83.36	80.22	81.76
Ours	**97.12**	**90.40**	**89.18**	**83.93**	**86.12**

4.3 Comparison with the State of the Art

When designating pose or parsing as the main task, we compare our results with the state-of-the-art approaches in these fields. Considering that human body edge detection does not have benchmarking results on the datasets included in this paper, so edge detection is not involved this section.

Human parsing. We evaluate our model on two single-human parsing datasets: LIP and ATR. As shown in Table 2, when DeepLabV2 is deployed as the backbone, the proposed model achieves 56.08% mIoU. When HRNetV2-W48 is the backbone, the mIoU of HTCorrM is 56.85%, which consistently outperforms competitive methods. Particularly on *sock*, *glove* and *leg*, we observe approximately 5% improvement in per-class IoU. MuLA [14], JPPNet [60] and CE2P [1] are three methods that aggregate only one related cue with parsing, while our model explores the correlation among all the three factors,and improves the performance by 7.55% mIoU, 5.48% mIoU and 3.75% mIoU, respectively. "Ours (HRNetV2-W48)" brings 0.95% and 0.51% performance gain to HRNetV2 [55] (55.90%) and GRN [67] (56.34%), respectively. Meanwhile, we implement HRNetV2 [55] with a smaller input size (384×384 *vs.* 473×473) and we do not utilize the flip-testing augmentation strategy. So the performance is slightly lower than reported. "Ours (HRNetV2-W48)" improves the accuracy of

0162-8828 (c) 2021 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.
Authorized licensed use limited to: Peking University. Downloaded on January 12,2022 at 15:26:18 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2021.3108771, IEEE Transactions on Pattern Analysis and Machine Intelligence

TABLE 4
Method comparison on the multi-human parsing dataset CIHP. mIoU (%) is reported. DeepLabV2 with the ResNet101 architecture is the task-specific backbone. HRNetV2-W48 is the unified backbone. We observe that our method is very competitive compared with the state of the art and that the two backbones have very close accuracy.

Method	Year	Backbone	Input size	mIoU
PGN [45]	2018	ResNet101	512×512	55.80
Graphonomy [6]	2019	DeepLabV3+	512×512	58.58
M-CE2P [1]	2019	ResNet101	384×384	59.50
Parsing R-CNN [72]	2019	ResNeXt101	512×864	59.80
CorrPM [20]	2020	ResNet101	384×384	60.18
RP R-CNN [73]	2020	ResNet50	512×1400	60.20
Grapy-ML [74]	2019	Xception	512×512	60.60
BraidNet [75]	2019	ResNet101	384×384	60.62
SNT [76]	2019	ResNet101	473×473	60.87
PCNet [68]	2020	ResNet101	512×512	61.05
Ours		ResNet101	384×384	**61.83**
Ours		HRNetV2-W48	384×384	**61.99**

TABLE 5
Method comparison on the MPII validation set for pose estimation. Our model achieves 90.7% in Mean PCKh and the highest score in almost all the joints. Especially, for *elbow* and *wrist* which are prone to be occluded, we improve their PCKh by 1.0% and 0.6%, respectively.

Method	Head	Sho.	Elb.	Wri.	Hip	Knee	Ank.	Mean
Newell *et al.* [3]	96.5	96.0	90.3	85.4	88.8	85.0	81.9	89.2
Yang *et al.* [34]	96.8	96.0	90.4	86.0	89.5	85.2	82.3	89.6
Tang *et al.* [77]	95.6	95.9	90.7	86.5	**89.9**	86.6	82.5	89.8
SimpleBaseline [2]	97.0	95.9	90.3	85.0	89.2	85.3	81.3	89.6
HRNet [36]	**97.1**	95.9	90.3	86.4	89.1	87.1	**83.3**	90.3
Ours	**97.1**	**96.4**	**91.3**	**87.0**	89.6	**87.2**	**83.3**	**90.7**

implemented HRNetV2 from 53.48% mIoU to 56.85% mIoU (+3.37% mIoU). In fact, PCNet [68], CNIF [69] and HHP [70] and QANet [71] use larger input size (*e.g.*, 473×473 or 512×384) and test-time augmentation strategies (*e.g.*, multi-scale testing and flip testing). Moreover, some graph-based methods need to carefully design the adjacency of nodes (body parts) according to each dataset but our approach does not require this step. Although these methods report higher accuracy (maximally by 2.7%), they are more complex than our method and potentially complementary with ours. Table 3 reports the results on ATR with the recent approaches. The proposed method brings a significant performance gain on each metric. Particularly, our model attains 4.36% boost for F-1 score compared with the prior state-of-art method TGPNet, and Precision is improved from 83.36% to 89.18%.

Experimental results on multi-human parsing dataset CIHP are reported in Table 4. With the same or even smaller input size, our model achieves 61.99% mIoU with the unified backbone HRNetV2-W48 and 61.83% mIoU with the specific backbone DeepLabV2. Both mark better performance than the state-of-the-art methods. Compared with PGN [45] which also uses edge cue for human parsing task, our model improves the accuracy by 6.19%mIoU. HTCorrM outperforms RP R-CNN [73] by 1.79% mIoU. Graphonomy [6], Parsing R-CNN [72], and Grapy-ML [74] adopt more complex models as the backbone, and the size of their input images (512×512, 512×864 and 512×512) is larger than ours (384×384). In comparison, the proposed HTCorrM yields 3.41%, 2.19% and 1.39% improvement in terms of mIoU, respectively.

Pose estimation. Table 5 and Table 6 provide the comparison with several pose estimation methods on MPII and LIP, respectively. Benefit from parsing and edge cues, our approach achieves 90.7% PCKh. It is worth noting that our model get the highest score in almost all the joints. Since the parsing annotations are not offered in the original dataset and they are generated by current human parsing algorithm, our model is flexible and has a low-complexity to be deployed with no manual labeling cost. We also conduct pose estimation experiments on LIP. MuLA and PIL both leverage the parsing factor to help pose estimation. But they do not uncover the correlation between the two tasks. Meanwhile, they adopt Hourglass as the backbone, which has the similar complexity with our backbone HRNetV2-W48 but is difficult to converge. Our PCKh is 87% with a margin of 1.4% PCKh over PIL and 1.6% PCKh over MuLA.

TABLE 6
Method comparison on the LIP validation set for pose estimation. We adopt HRNetV2-W48 as backbone, whose GFLOPs are almost the same as Hourglass. Our model reports higher PCKh than state-of-the-art methods such as PIL, MuLA and GCM.

Method	Backbone	PCKh
PIL [5]	VGG16	75.0
MuLA [14]	VGG16	76.0
GCM [78]	MobileNet	84.0
MuLA [14]	Hourglass	85.4
PIL [5]	Hourglass	85.6
Ours	HRNetV2-W48	**87.0**

Visualization Results. Fig. 8 represents the quality results on two human parsing datasets CIHP and LIP with the specific backbone DeepLabV2 and unified backbone HRNetV2-W48. We compare our model with three current methods: DeepLabV2 [16], CE2P [1] and CorrPM [20]. We observe from the second row in (a) that by aggregating with pose information, our model can learn the global body structure of human, thereby accurately segments the region of *legs* and *shoes*, which are wrongly predicted by DeepLabV2 and CE2P. Also with the help of edge information, our framework successfully locates the semantic boundary of *pants* and *upper-clothes* shown in the first row of (b). Compared with CorrPM, the end-to-end trainable model HTCorrM digs the correlation between each person on image level instead of a single person on mask level. Hence, in the second row of (a), it captures the interaction between people and correctly distinguishes *left arm* and *right arm* of the two people holding hands.

4.4 Analysis

In this section, we compare the performance of correlation with that of self-attention module, multi-task learning and feature concatenation. Computational complexity and different backbones are discussed. We visualize the relation map of three tasks, and the output of pose and edge. We also analyze the impact of the "duplicated" feature, and the impact of the auxiliary task's accuracy to the main task, *etc.* The experiments are performed over human parsing on CIHP in Table 7, over pose estimation on MPII and LIP in Table 8, and over human edge detection on LIP in Table 9.

Correlation vs. self-attention module. To compare correlation with self-attention module, firstly, we train a baseline model which only employs the backbone to perform

0162-8828 (c) 2021 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.
Authorized licensed use limited to: Peking University. Downloaded on January 12,2022 at 15:26:18 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2021.3108771, IEEE Transactions on Pattern Analysis and Machine Intelligence

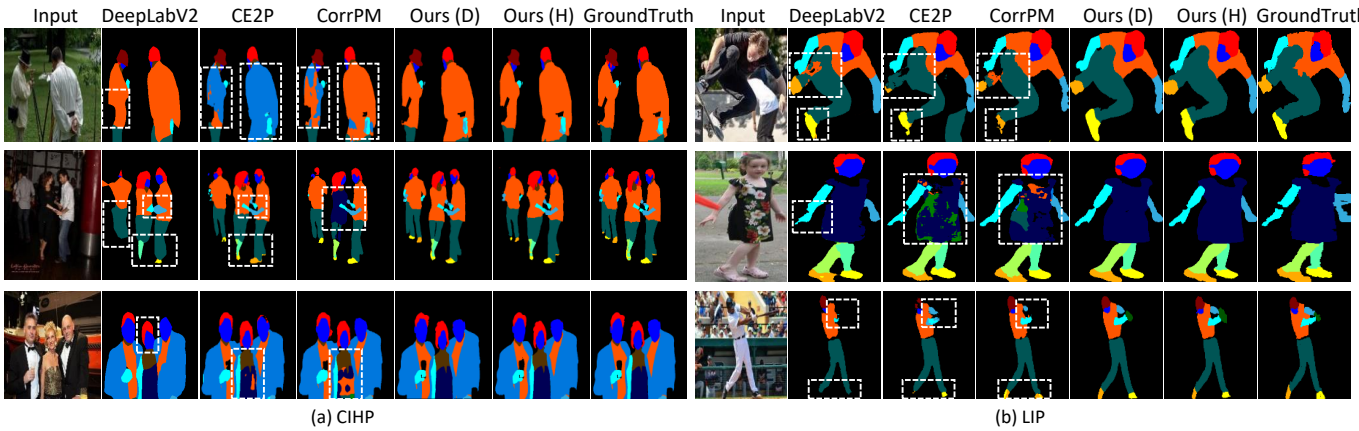


Fig. 8. Qualitative comparisons of DeepLabV2 [16], CE2P [1], CorrPM [20] and the proposed HTCorrM (extension to [20]) on (a) multi-human parsing dataset dataset CIHP, and (b) single-human parsing dataset LIP. "Ours (D)" and "Ours (H)" are models using DeepLabV2 and HRNetV2-W48 as the backbone, respectively. In (a), our model is precise on clothes by exploring relationship over the entire image to distinguish clothes between different individuals. In (b), HTCorrM eliminates parsing errors (*e.g.*, inaccurate boundaries between upper-clothes and pants in the first row). Some poor predictions are shown in white rectangles.

TABLE 7
Variant comparison on the multi-human parsing dataset CIHP. We adopt DeepLabV2 as baseline, which achieves 50.58% mIoU. "SA": self-attention module; "MTL": multi-task learning; "Concat": feature concatenation; "Correlation": the proposed correlation strategy; "✓" stands for selecting one feature and "✓✓" means fusing the same two features. "†" means pose annotations are generated by [56], and the rest of the experiments use [2] as the pose labels.

Method	Strategy	Feature			mIoU
		Parsing	Edge	Pose	
1	SA	✓	-	-	51.80
2	MTL	✓	✓	✓	52.29
3	Concat	✓	✓		54.53
4		✓		✓	55.20
5		✓	✓	✓	56.32
6	Correlation	✓	✓		59.69
7		✓		✓	59.59
8		✓	✓✓		60.14
9		✓		✓✓	60.02
10		✓	✓	✓	61.77†
11		✓	✓	✓	**61.83**

human parsing task in Table 7. It achieves 50.58% mIoU. Then we add a self-attention module to it (implemented by the standard non-local network), resulting in "Method 1" in Table 7. This model improves the mIoU of baseline by 1.22%, while our model ("Method 11") boosts the mIoU by 11.25%. It shows that solely paying attention to the parsing feature is not enough, accordingly, the proposed HNL aggregates multiple features heterogeneously and exploits the among-task correlation.

Correlation vs. multi-task learning. We remove HNL from HTCorrM, resulting in "Method 2" as a way of multi-task learning in Table 7. We take the prediction of parsing encoder as the output. Next we calculate the accuracy between it and the parsing groundtruth. This model obtains 1.71% gain in terms of mIoU compared with baseline, which illustrates that simultaneously performing the three tasks can bring improvement. However, the network does not explore the correlation among the three, so the performance gain is quite limited compared with our 11.25 points.

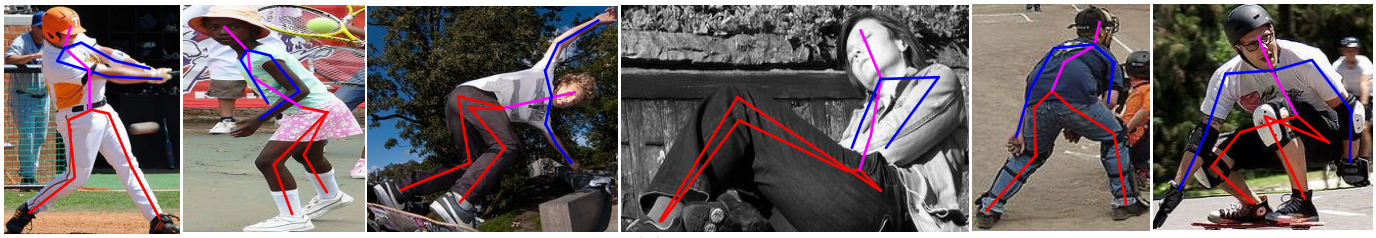
Correlation vs. concatenation. To validate the effectiveness of the correlation strategy in HNL, we compare it with concatenation methods. From Table 7, "Method 3" boosts 3.95% mIoU to the baseline by concatenating parsing with edge feature, and "Method 4" concatenates parsing with pose feature, boosting 4.62% mIoU to the baseline. It shows pose and edge are both beneficial for parsing task. Specifically, after concatenating all the three features, the performance of "Method 5" is 5.74% higher than Baseline. Then we conduct correlation experiments between parsing and pose factor, generating "Method 7", and between parsing and edge factor, generating model "Method 6". This two model bring 4.39% and 5.16% mIoU gain to "Method 4" and "Method 3". The improvement of correlation is more significant compared with baseline model, yielding 9.01% and 9.11% increases in terms of mIoU. There are even 3.27% mIoU gain compared with "Method 5". Moreover, when all the three representations are correlated together, the mIoU reaches the highest 61.83%. It demonstrates that our correlation fusion strategy is more superior and effective than the concatenation approach.

Does improvement come from the increased number of channels or the aggregation of multiple tasks? In HNL, we aggregate three feature maps from different tasks along the channel dimension, which will triple the channel number. To study whether the gain is from the increased number of channels or from the integration of multiple tasks, we concatenate two edge (or pose) feature maps with one parsing feature along channel dimension, demonstrated as "Method 8" (or "Method 9") in Table 7. They have the same channel number as the proposed model, while the mIoU is about 1.8% lower than it. It shows that the improvement is from the aggregation of parsing, pose and edge cues. None of the three is dispensable. This also proves that the correlated cue provided by any one of the three is irreplaceable, so it is necessary to fuse all of them in a unified model.

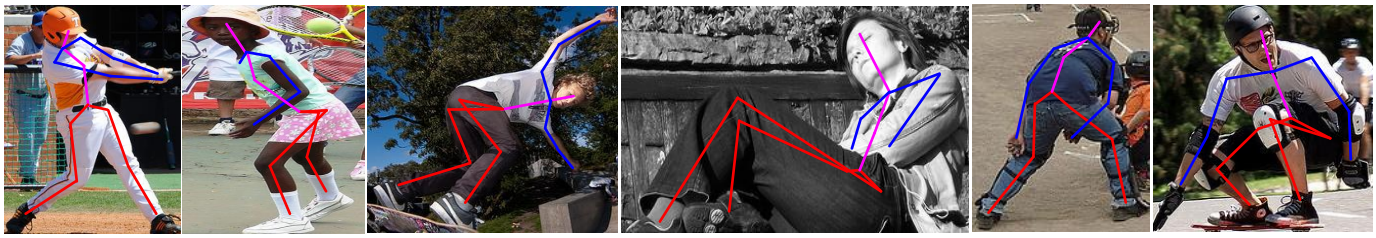
Impact of using different pose estimators to generate groundtruth. In Table 7, we also explore the influence of different pose annotations to parsing performance. Utilizing the pose estimator [2] to generate the pose label, our model achieves the highest 61.83% mIoU. When we change to alphapose [56] whose performance is lower than [2] on public pose estimation dataset COCO [59], the performance

0162-8828 (c) 2021 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.
Authorized licensed use limited to: Peking University. Downloaded on January 12,2022 at 15:26:18 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2021.3108771, IEEE Transactions on Pattern Analysis and Machine Intelligence



(a) Baseline



(b) Ours

Fig. 9. Qualitative pose estimation predictions on LIP. These examples contain occlusions, various human poses and cluster backgrounds. From the results of baseline model and ours, we observe that parsing and edge are beneficial for pose estimation. For instance, in column 5, the right wrist is missing in the baseline since the right arm is occluded by the leg. In comparison, parsing and edge feature provide contextual cues for the system to find this hard point.

TABLE 8
Analysis of computational complexity *vs.* accuracy on MPII and LIP. We use training / inference time (seconds per iteration), number of parameters and GFLOPs to measure complexity, and PCKh for accuracy. Two backbones (baselines) and the state-of-the-art method MuLA [14] are compared. The main conclusion is that our method adds marginal computational overheads to the baselines and has lower complexity than MuLA. "†" means removing "duplicated" features in HNL.

Method	Backbone	Feature			Train sec./iter.	Infer sec./iter.	#Params	GFLOPs	MPII	LIP
		Pose	Parsing	Edge						
MuLA	Hourglass	✓	✓		1.967	0.183	46.2M	46.3	-	85.4
Ours	HRNetV2-W32	✓			0.515	0.054	28.5M	9.5	90.30	85.38
		✓	✓		0.532	0.056	28.5M	9.6	90.32	86.21
		✓	✓	✓	0.557	0.060	28.5M	9.7	**90.34**	**86.76**
Ours	HRNetV2-W48	✓			0.607	0.059	63.6M	19.6	90.40	86.11
		✓	✓		0.627	0.063	63.6M	19.6	90.66	86.62
		✓	✓	✓	0.632	0.065	63.7M	19.9	90.69†	86.87†
		✓	✓	✓	0.633	0.065	63.7M	19.9	**90.73**	**87.00**

is 61.77%. This result shows our model is robust to the pose labels and has low deployment complexity with almost no performance drop.

Parsing, pose and edge can all be improved when being implemented as the main task. As shown in Table 7, "Method 7", "Method 6" and "Method 11" achieve 59.59%, 59.69% and 61.83% in terms of mIoU. Compared with 50.58% mIoU of parsing baseline, this result indicates that the parsing performance is improved when specified as the main task. Since the edge cue helps to distinguish the border between two parts, and the pose cue can assist parsing to perceive a reasonable body structure, correlating the three of them can yield performance gain.

Meanwhile for the pose estimation task, we report the experimental results in Table 8 on two datasets: MPII and LIP. HRNetV2-W32 and HRNetV2-W48 are applied as two baselines. The performance on MPII is improved when correlating pose with parsing or edge feature compared with the baseline model, and there are more increase on the larger model HRNetV2-W48 than HRNetV2-W32. On LIP, when correlating pose with parsing feature only, the PCKh achieves 86.21%, which has 0.83% gain compared with the baseline model HRNetV2-W32. Furthermore, if we correlate pose with both parsing and edge factors, the model achieves 86.76%, which even outperforms baseline HRNetV2-W48 by 0.65%. HRNetV2-W48 has more than twice as many parameters as our model and has much higher GFLOPs than our network. When changing the backbone to HRNetV2-W48, our model obtains 0.89% performance gain, achieving a PCKh of 87.00%, while the parameters and GFLOPs of this two models are almost the same.

Qualitative pose estimation examples on LIP are given in Fig. 9. This dataset contains samples with various human postures, complex background and rapidly changing appearance. Such patterns usually lead to poor keypoint location identifications. We depict the predictions of the HRNetV2-W48 baseline and ours. It is shown that our model presents more precise predictions than the baseline despite those distracting factors. In Column 3 of Fig. 9, the baseline model wrongly predicts the left hip to the right hip. In Column 5, the proposed model accurately locates the right wrist which is missing in the baseline.

Table 9 provides the result of human edge detection. We deploy two backbones: the unified backbone HRNetV2-W48 [55] and the task-specific backbone BDCN [13]. The baseline model HRNetV2-W48 and BDCN obtain 0.649 and 0.656 ODS. After we add parsing to the baseline model and seek the correlation between them, the two backbones respec-

0162-8828 (c) 2021 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.
Authorized licensed use limited to: Peking University. Downloaded on January 12,2022 at 15:26:18 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2021.3108771, IEEE Transactions on Pattern Analysis and Machine Intelligence

TABLE 9
Evaluation of the benefit of parsing and pose to edge detection on the LIP validation set. ODS and OIS are used as metrics. We adopt BDCN [13] and HRNetV2-W48 [55] as backbones, respectively. By correlating edge with parsing and pose, the performance of edge detection is improved.

Backbone	Feature			ODS	OIS
	Edge	Parsing	Pose		
HRNetV2-W48	✓			0.649	0.654
	✓	✓		0.658	0.673
	✓	✓	✓	0.674	0.686
BDCN	✓			0.656	0.668
	✓	✓		0.671	0.682
	✓	✓	✓	**0.684**	**0.693**

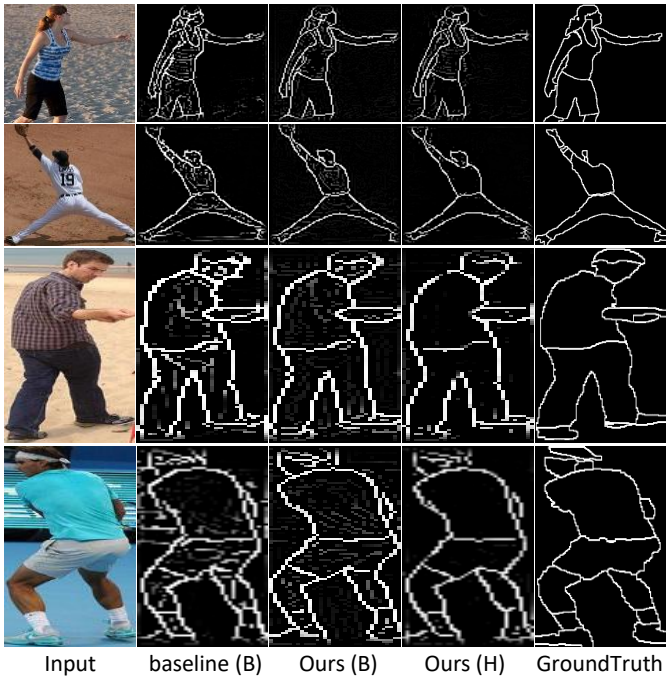


Fig. 10. Sample edge detection results of baseline BDCN "baseline (B)" and ours on LIP. "Ours (B)" and "Ours (H)" are models using BDCN and HRNetV2-W48 as the backbone, respectively. "baseline (B)" tends to detect very minor edges including stripes on upper-clothes in the top row and wrinkles on pants in the bottom row. We observe that our models ignore the very detailed and noisy textures and thus have more desirable human boundaries.

tively lead to 0.9% and 1.5% ODS improvement compared with the baseline. When both parsing and pose features are fused with edge representations, the accuracy gets even higher: improvement of the two backbones over the baseline is 2.5% and 2.8% when using BDCN and HRNetV2-W48 backbones, respectively.

We also show the qualitative edge detection results of the backbone BDCN, our model using BDCN backbone and our model using HRNetV2-W48 backbone on the LIP dataset. As shown in Fig. 10, there are many false positives in the prediction of baseline model. For instance, in the first and last column, cross stripes on upper clothes and wrinkles on pants are misleadingly regarded as edges. Our model has the ability to eliminate the false positives and the results are more precise.

All above quantitative and qualitative results validate the superiority and effectiveness of HTCorrM. It shows that parsing, pose and edge are associated and each of them can provide complementary cue to refine the main task and strengthen the model.

A unified backbone *vs.* task-specific backbones. We compare the unified backbone HRNetV2-W48 and task-specific backbones on two human parsing datasets LIP and CIHP in Table 2 and Table 4, respectively, and on edge detection dataset LIP in Table 9. On LIP, "Ours (HRNetV2-W48)" achieves 56.85% mIoU and slightly outperforms "Ours (DeepLabV2)" by 0.77% mIoU. On CIHP, the mIoU of HRNetV2-W48 is higher than that of DeepLabV2 by 0.16%. For edge detection, on the other hand, using BDCN as backbone outperforms the HRNetV2-W48 backbone by 1.0% in ODS and 0.7% in OIS. Overall, the two types of backbones are both effective and achieve comparable accuracy. It further confirms that our system can accommodate different backbones.

Computational complexity. In Table 8, we analyze the computational complexity of the baseline (HRNetV2-W32 and HRNetV2-W48), our method HTCorrM and MuLA [14] in terms of training / inference time, number of parameters, GFLOPs, and PCKh on MPII and LIP datasets. The training experiments are conducted on 2 TITAN V GPU cards, with the batch size of 32. For inference, a single TITAN V GPU card is used and the batch size is 1. The comparisons indicate the following speculations. First, HTCorrM does not add significantly more parameters and higher computational cost compared with the baseline. For example, when HRNetV2-W48 is adopted as the baseline, the number of parameters of our system is only 0.1% higher than the baseline. Besides, the training and inference time is increased from 0.607s to 0.633s and from 0.054s to 0.060s, respectively. The above-mentioned increase in complexity is relatively low if we consider that our method is higher than the baseline by 0.33% and 0.89% PCKh on MPII and LIP, respectively.

Second, when compared with MuLA, our method is higher in accuracy and yet lower in computational complexity. MuLA is a multi-task learning framework, generating pose and parsing predictions simultaneously. In the comparison, our method and MuLA use 19.9 GFLOPs and 46.3 GFLOPs, respectively. In fact, the pose and parsing encoders in MuLA take nearly half of the computation ($\frac{20.9GFLOPs}{46.3GFLOPs}$ = 45.1%), while the encoders and HNL in our method only take less than 1.5% of the total GFLOPS. Moreover, MuLA consumes 1.967s and 0.183s in the training and inference, respectively, and our method is 67.8% (1.334s to 1.967s) and 64.6% (0.118s to 0.183s) lower. The above analysis demonstrates the characteristics of our method: by focusing on a main task instead of multiple tasks, our model adds marginal parameters and computational complexity to the baseline due to the light-weight encoders and HNL, and achieves a higher accuracy.

Impact of the "duplicated" feature in HNL. HNL has three input features, and one of them is the same with the main task feature. To investigate the impact of this "duplicated" feature, we remove it from the input and only deliver the other two features into the HNL module. Experiments are conducted on the pose estimation task, and results are shown in Table 8. Comparing with the full model (HRNetV2-W48 as the backbone), removing the "duplicated" feature causes a slight 0.04% and 0.13% drop in mIoU on MPII and LIP, respectively (the computation

0162-8828 (c) 2021 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.
Authorized licensed use limited to: Peking University. Downloaded on January 12,2022 at 15:26:18 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2021.3108771, IEEE Transactions on Pattern Analysis and Machine Intelligence

TABLE 10
Impact of auxiliary task (pose and edge) accuracy on the main task (parsing) on LIP. (a): we present the pose estimation accuracy (PCKh, %) and parsing accuracy (mIoU, %) obtained by two pose encoders structures (for details see Section 3.2). (b): we use two edge decoders (for details see Section 3.2), and obtain different edge detection accuracy (ODS, OIS) and parsing accuracy (mIoU, %). The edge encoder in (a) and the pose encoder in (b) are fixed, as described in Section 3.2. We use the unified backbone HRNetV2-W48. When the accuracy of pose and edge auxiliary tasks is increased, we observe improvement of the main task.

(a) Impact of pose encoders on pose and parsing accuracy

Strategy	encoder	PCKh	mIoU
Correlation	conv×2	45.3	54.48
Correlation	deconv×2	**56.9**	**56.85**

(b) Impact of edge encoders on edge and parsing accuracy

Strategy	encoder	ODS	OIS	mIoU
Correlation	conv×1	0.528	0.531	53.29
Correlation	conv×3	**0.630**	**0.635**	**56.85**

complexity remains mostly the same). In fact, having this feature in HNL enables a self-attention process with the main task feature. So essentially HNL not only explores cross-attention between different modalities (*e.g.*, pose and parsing, pose and edge), but also has self-attention within the same modality (*e.g.*, pose and pose), both of which help to reach the best accuracy. That said, we note that the self-attention mechanism is less important than cross attention, because the latter utilizes the strong complementary nature between tasks. The above result and analysis indicate the input feature which is the same with the main task feature is not necessarily "duplicated".

Higher accuracy of the auxiliary task helps improve the main task. We analyze the impact of the accuracy of each auxiliary task (i.e., pose or edge) on the main task (parsing) on the LIP dataset, by modifying the structure of one auxiliary task's encoder while fixing the other. For the pose encoder, we replace the two deconvolution layers with two convolution layers. For the edge encoder, we only utilize one convolution layer, which takes features from G_3 as the input. Results are summarized in Table 10. We observe that a higher accuracy of the auxiliary task (pose or edge) helps improve performance of the main task (parsing). For example, compared with the edge encoder with only one convolutional layer, a stronger encoder (with three convolutional layers) improves the ODS from 0.528 to 0.630; at the same time, the main task accuracy is improved from 53.29% to 56.85% (+3.56%).

Visualization of the relation maps on the three tasks. We visualize the relation maps of human parsing, pose estimation and edge detection when they are specified as the main task. As shown in Fig. 11, we observe that higher responses (in red) on the relation maps (bottom row) are in areas highly correlated with the query points (top row). For example, the high response in Column 3 spreads around the keypoint *wrist*, and the lower responses (in blue) are seen at areas less relevant to the query. These results show that when each of the three is designated as the main task, the proposed HTCorrM can discover correlated pixels around for each query, leading to its representations that consider more contextual cues.

Fig. 11. Visualization of query points and relation maps of three tasks: (a) human parsing, (b) pose estimation and (c) edge detection. In the first row, the red point in each image is the query point. From left to right, the query point is: the point on *pants*, point on *sock*, keypoint *wrist*, keypoint *shoulder*, point on the border of the man's *arm*, point on the border of the woman's *skirt*. Their relation maps are shown right below in the second row. Areas marked in red contain high-response pixels that have the same semantics with the query, while blue areas indicate low responses. Note that in (c), relation maps are not merged with the original image for better visualization. The results showcase that the proposed HTCorrM can discover correlated pixels around for each query, leading to its representations that consider more contextual cues.

5 Conclusion

In this paper, we propose a Human Task Correlation Machine (HTCorrM) to explore the relationship among parsing, semantic edge and body keypoint features. When specified as the main task, the performance of any one the three tasks can be improved by correlating it with the other two auxiliary tasks. With the Heterogeneous Non-Local (HNL) module, the proposed model utilizes the complementary cues of above three factors through feature correlation which is superior to concatenation. HNL does not add much computation complexity to the backbone and gives comparable accuracy. The whole model explores the correlation on the image level instead of the mask level, and is end-to-end learnable. Experiments on four benchmarks demonstrate the effectiveness of the proposed method. Moreover, the requirement of manual labels of our system is not significantly heavy, since it can leverage the existing pose or parsing algorithms to generate the annotations to supervise the model.

Future work. We will go further on studying the relationships of the existing three tasks with other complementary tasks of human, such as action recognition and attribute recognition. Since the HNL has the potential to integrate the relevant cues into the hybrid representation with few computation cost increased. The corresponding encoder is to be designed to better fit the feature learning and model the correlations among all the tasks.

HTCorrM can be extended to other vision tasks which are also associated. It could be face analysis such as face detection, landmark localization and gender recognition; or depth-related tasks such as depth estimation and point cloud segmentation; or low-level area such as image inpainting, super resolution and image denoising. While these applications are mostly studied in the multi-task fashion, we will apply our method to explore how to benefit a main task utilizing correlations among different tasks.

0162-8828 (c) 2021 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.
Authorized licensed use limited to: Peking University. Downloaded on January 12,2022 at 15:26:18 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2021.3108771, IEEE Transactions on Pattern Analysis and Machine Intelligence

REFERENCES

[1] T. Ruan, T. Liu, Z. Huang, Y. Wei, S. Wei, and Y. Zhao, "Devil in the details: Towards accurate single and multiple human parsing," in *AAAI*, vol. 33, 2019, pp. 4814–4821.

[2] B. Xiao, H. Wu, and Y. Wei, "Simple baselines for human pose estimation and tracking," in *ECCV*, 2018, pp. 466–481.

[3] A. Newell, K. Yang, and J. Deng, "Stacked hourglass networks for human pose estimation," in *ECCV*. Springer, 2016, pp. 483–499.

[4] J. Dai, K. He, and J. Sun, "Instance-aware semantic segmentation via multi-task network cascades," in *CVPR*, 2016, pp. 3150–3158.

[5] X. Nie, J. Feng, Y. Zuo, and S. Yan, "Human pose estimation with parsing induced learner," in *CVPR*, 2018, pp. 2100–2108.

[6] K. Gong, Y. Gao, X. Liang, X. Shen, M. Wang, and L. Lin, "Graphonomy: Universal human parsing via graph transfer learning," in *CVPR*, 2019, pp. 7450–7459.

[7] Y. Lu, K. Boukharouba, J. Boonært, A. Fleury, and S. Lecoeuche, "Application of an incremental svm algorithm for on-line human recognition from video surveillance using texture and color features," in *Neurocomputing*, vol. 126. Elsevier, 2014, pp. 132–140.

[8] Y. Wang, D. Tran, Z. Liao, and D. Forsyth, "Discriminative hierarchical part-based models for human parsing and action recognition," vol. 13, no. Oct, 2012, pp. 3075–3102.

[9] C. Wang, Y. Wang, and A. L. Yuille, "An approach to pose-based action recognition," in *CVPR*, 2013, pp. 915–922.

[10] C. Ma, Y. Li, F. Yang, Z. Zhang, Y. Zhuang, H. Jia, and X. Xie, "Deep association: End-to-end graph-based learning for multiple object tracking with conv-graph neural network," in *ICMR*, 2019, pp. 253–261.

[11] J. H. Yoon, C. Lee, M. Yang, and K. Yoon, "Online multi-object tracking via structural constraint event aggregation," in *CVPR*, 2016, pp. 1392–1400.

[12] Y. Luo, Z. Zheng, Z. Liang, G. Tao, J. Yu, and Y. Yi, "Macro-micro adversarial network for human parsing," in *ECCV*, 2018.

[13] J. He, S. Zhang, M. Yang, Y. Shan, and T. Huang, "Bdcn: Bi-directional cascade network for perceptual edge detection," in *TPAMI*, 2020, pp. 1–1.

[14] X. Nie, J. Feng, and S. Yan, "Mutual learning to adapt for joint human parsing and pose estimation," in *ECCV*, 2018, pp. 502–517.

[15] F. Xia, P. Wang, X. Chen, and A. L. Yuille, "Joint multi-person pose estimation and semantic part segmentation," in *CVPR*, 2017, pp. 6769–6778.

[16] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs," in *TPAMI*, vol. 40, no. 4. IEEE, 2017, pp. 834–848.

[17] K. Gong, X. Liang, D. Zhang, X. Shen, and L. Lin, "Look into person: Self-supervised structure-sensitive learning and a new benchmark for human parsing," in *CVPR*, 2017, pp. 932–940.

[18] L. Ladicky, P. H. Torr, and A. Zisserman, "Human pose estimation using a joint pixel-wise and part-wise formulation," in *CVPR*, 2013, pp. 3578–3585.

[19] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask r-cnn," in *ICCV*, 2017, pp. 2961–2969.

[20] Z. Zhang, C. Su, L. Zheng, and X. Xie, "Correlating edge, pose with parsing," in *CVPR*, June 2020.

[21] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *CVPR*, 2015, pp. 3431–3440.

[22] L.-C. Chen, Y. Yang, J. Wang, W. Xu, and A. L. Yuille, "Attention to scale: Scale-aware semantic image segmentation," in *CVPR*, 2016, pp. 3640–3649.

[23] J. Zhao, J. Li, X. Nie, F. Zhao, Y. Chen, Z. Wang, J. Feng, and S. Yan, "Self-supervised neural aggregation networks for human parsing," in *CVPRW*, 2017, pp. 7–15.

[24] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical image computing and computer-assisted intervention*, 2015, pp. 234–241.

[25] V. Badrinarayanan, A. Handa, and R. Cipolla, "Segnet: A deep convolutional encoder-decoder architecture for robust semantic pixel-wise labelling," in *Computer Science*, 2015.

[26] X. Peng, R. S. Feris, X. Wang, and D. N. Metaxas, "A recurrent encoder-decoder network for sequential face alignment," 2016, pp. 38–56.

[27] H. Noh, S. Hong, and B. Han, "Learning deconvolution network for semantic segmentation," in *ICCV*, 2015.

[28] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in *CVPR*, 2017, pp. 2881–2890.

[29] X. Liang, C. Xu, X. Shen, J. Yang, S. Liu, J. Tang, L. Lin, and S. Yan, "Human parsing with contextualized convolutional neural network," in *ICCV*, 2015, pp. 1386–1394.

[30] J. Li, J. Zhao, Y. Wei, C. Lang, Y. Li, T. Sim, S. Yan, and J. Feng, "Multiple-human parsing in the wild," in *arXiv preprint arXiv:1705.07206*, 2017.

[31] Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh, "Realtime multi-person 2d pose estimation using part affinity fields," in *CVPR*, 2017, pp. 7291–7299.

[32] G. Papandreou, T. Zhu, N. Kanazawa, A. Toshev, J. Tompson, C. Bregler, and K. Murphy, "Towards accurate multi-person pose estimation in the wild," in *CVPR*, 2017, pp. 4903–4911.

[33] A. Newell, Z. Huang, and J. Deng, "Associative embedding: End-to-end learning for joint detection and grouping," in *NeurIPS*, 2017, pp. 2277–2287.

[34] W. Yang, S. Li, W. Ouyang, H. Li, and X. Wang, "Learning feature pyramids for human pose estimation," in *ICCV*, 2017, pp. 1281–1290.

[35] S.-E. Wei, V. Ramakrishna, T. Kanade, and Y. Sheikh, "Convolutional pose machines," in *CVPR*, 2016, pp. 4724–4732.

[36] K. Sun, B. Xiao, D. Liu, and J. Wang, "Deep high-resolution representation learning for human pose estimation," in *CVPR*, 2019, pp. 5693–5703.

[37] J. Kittler, "On the accuracy of the sobel edge detector," in *Image & Vision Computing*, vol. 1, no. 1, 1983, pp. 37–42.

[38] J. Canny, "A computational approach to edge detection," in *TPAMI*, vol. 8, no. 6, 1986, pp. 679–698.

[39] D. Martin, C. C. Fowlkes, and J. Malik, "Learning to detect natural image boundaries using local brightness, color, and texture cues," in *TPAMI*, vol. 26, no. 5, 2004, pp. 530–549.

[40] S. Xie and Z. Tu, "Holistically-nested edge detection," 2015, pp. 1395–1403.

[41] Z. Yu, C. Feng, M. Liu, and S. Ramalingam, "Casenet: Deep category-aware semantic edge detection," 2017, pp. 1761–1770.

[42] H.-S. Fang, G. Lu, X. Fang, J. Xie, Y.-W. Tai, and C. Lu, "Weakly and semi supervised human body part parsing via pose-guided knowledge transfer," in *arXiv preprint arXiv:1805.04310*, 2018.

[43] D. Jian, C. Qiang, X. Shen, J. Yang, and S. Yan, "Towards unified human parsing and pose estimation," in *CVPR*, 2014.

[44] L.-C. Chen, J. T. Barron, G. Papandreou, K. Murphy, and A. L. Yuille, "Semantic image segmentation with task-specific edge detection using cnns and a discriminatively trained domain transform," in *CVPR*, 2016, pp. 4545–4554.

[45] K. Gong, X. Liang, Y. Li, Y. Chen, M. Yang, and L. Lin, "Instance-level human parsing via part grouping network," in *ECCV*, 2018, pp. 770–785.

[46] Z. Hu, M. Zhen, X. Bai, H. Fu, and C.-l. Tai, "Jsenet: Joint semantic segmentation and edge detection network for 3d point clouds," in *arXiv preprint arXiv:2007.06888*, 2020.

[47] X. Wang, R. Girshick, A. Gupta, and K. He, "Non-local neural networks," in *CVPR*, 2018, pp. 7794–7803.

[48] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," 2017, pp. 5998–6008.

[49] J. Jia, "Psanet: Point-wise spatial attention network for scene parsing," in *ECCV*, 2018.

[50] J. Fu, J. Liu, H. Tian, Y. Li, Y. Bao, Z. Fang, and H. Lu, "Dual attention network for scene segmentation," in *CVPR*, 2018, pp. 3146–3154.

[51] Z. Huang, X. Wang, L. Huang, C. Huang, Y. Wei, and W. Liu, "Ccnet: Criss-cross attention for semantic segmentation," in *ICCV*, October 2019.

[52] Z. Zhu, M. Xu, S. Bai, T. Huang, and X. Bai, "Asymmetric non-local neural networks for semantic segmentation," in *ICCV*, October 2019.

[53] Y. Zhang, K. Li, K. Li, B. Zhong, and Y. Fu, "Residual non-local attention networks for image restoration," in *arXiv preprint arXiv:1903.10082*, 2019.

[54] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *CVPR*, 2016, pp. 770–778.

[55] J. Wang, K. Sun, T. Cheng, B. Jiang, C. Deng, Y. Zhao, D. Liu, Y. Mu, M. Tan, X. Wang, W. Liu, and B. Xiao, "Deep high-resolution representation learning for visual recognition," in *TPAMI*, 2019.

[56] H.-S. Fang, S. Xie, Y.-W. Tai, and C. Lu, "RMPE: Regional multi-person pose estimation," in *ICCV*, 2017.

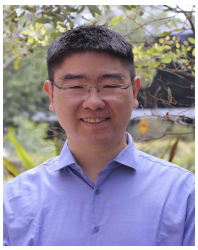
0162-8828 (c) 2021 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.
Authorized licensed use limited to: Peking University. Downloaded on January 12,2022 at 15:26:18 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2021.3108771, IEEE Transactions on Pattern Analysis and Machine Intelligence

[57] A. Oliver, A. Odena, C. Raffel, E. D. Cubuk, and I. J. Goodfellow, "Realistic evaluation of deep semi-supervised learning algorithms," in *arXiv preprint arXiv:1804.09170*, 2018.
[58] M. Andriluka, L. Pishchulin, P. V. Gehler, and B. Schiele, "2d human pose estimation: New benchmark and state of the art analysis," 2014, pp. 3686–3693.
[59] T. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollar, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *ECCV*, 2014, pp. 740–755.
[60] X. Liang, G. Ke, X. Shen, and L. Liang, "Look into person: Joint body parsing and pose estimation network and a new benchmark," in *TPAMI*, vol. PP, no. 99, 2018, pp. 1–1.
[61] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *CVPR*. Ieee, 2009, pp. 248–255.
[62] Y. Zhuang, F. Yang, L. Tao, C. Ma, Z. Zhang, Y. Li, H. Jia, X. Xie, and W. Gao, "Dense relation network: Learning consistent and context-aware representation for semantic image segmentation." in *International Conference on Image Processing*, 2018, pp. 3698–3702.
[63] K. Yamaguchi, M. H. Kiapour, L. E. Ortiz, and T. L. Berg, "Parsing clothing in fashion photographs," in *CVPR*. IEEE, 2012, pp. 3570–3577.
[64] K. Yamaguchi, M. H. Kiapour, and T. L. Berg, "Paper doll parsing: Retrieving similar styles to parse clothing items," in *ICCV*, 2013.
[65] X. Liang, S. Liu, X. Shen, J. Yang, L. Liu, J. Dong, L. Lin, and S. Yan, "Deep human parsing with active template regression," in *TPAMI*, vol. 37, no. 12, 2015, p. 2402.
[66] X. Luo, Z. Su, J. Guo, G. Zhang, and X. He, "Trusted guidance pyramid network for human parsing," in *ACM MM*, 2018, pp. 654–662.
[67] T. Li, Z. Liang, S. Zhao, J. Gong, and J. Shen, "Self-learning with rectification strategy for human parsing," in *CVPR*, 2020.
[68] X. Zhang, Y. Chen, B. Zhu, J. Wang, and M. Tang, "Part-aware context network for human parsing," in *CVPR*, June 2020.
[69] W. Wang, Z. Zhang, S. Qi, J. Shen, Y. Pang, and L. Shao, "Learning compositional neural information fusion for human parsing," in *ICCV*, 2019.
[70] W. Wang, H. Zhu, J. Dai, Y. Pang, J. Shen, and L. Shao, "Hierarchical human parsing with typed part-relation reasoning," in *CVPR*, 2020.
[71] L. Yang, Q. Song, Z. Wang, Z. Liu, S. Xu, and Z. Li, "Quality-aware network for human parsing," in *arXiv preprint arXiv:2103.05997*, 2021.
[72] L. Yang, Q. Song, Z. Wang, and M. Jiang, "Parsing r-cnn for instance-level human analysis," in *CVPR*, 2019, pp. 364–373.
[73] L. Yang, Q. Song, Z. Wang, M. Hu, and S. Xu, "Renovating parsing r-cnn for accurate multiple human parsing," in *ECCV 2020*, 2020.
[74] H. He, J. Zhang, Q. Zhang, and D. Tao, "Grapy-ml: Graph pyramid mutual learning for cross-dataset human parsing," in *AAAI*, 2019.
[75] X. Liu, M. Zhang, W. Liu, J. Song, and T. Mei, "Braidnet: Braiding semantics and details for accurate human parsing," 2019, pp. 338–346.
[76] R. Ji, D. Du, L. Zhang, L. Wen, Y. Wu, C. Zhao, F. Huang, and S. Lyu, "Learning semantic neural tree for human parsing," in *ECCV*, 2020.
[77] W. Tang, P. Yu, and Y. Wu, "Deeply learned compositional models for human pose estimation," in *ECCV*, 2018, pp. 190–206.
[78] D. Osokin, "Global context for convolutional pose machines." in *arXiv: Computer Vision and Pattern Recognition*, 2019.

Ziwei Zhang received the B.S. degree in electronic engineering from Xidian University, Xi'an, China, in 2016. She is currently pursuing the Ph.D. degree with the Department of Computer Science and Technology, Peking University, Beijing, China. Her research interests are computer vision and deep learning, with focus on semantic segmentation and human parsing.

Chi Su is currently a General Manager of Artificial Intelligence Product Center at Kingsoft Cloud, Beijing, China. He received the Ph.D. from the Department of Computer Science and Technology in Peking University, Beijing, China. His research interests include computer vision and machine learning, with focus on object detection, object tracking, and human identification and recognition.



Liang Zheng (SM'20) received the B.E. in life science and Ph.D. in electronic engineering from Tsinghua University, China, in 2010 and 2015, respectively. He was a Post-Doctoral Researcher with Centre for Artificial Intelligence, University of Technology Sydney, Australia. He is now a Senior Lecturer in the School of Computing, Australian National University. He holds the CS Futures Fellowship and ARC DECRA Fellowship. His research interests include image retrieval, data synthesis, and object re-identification. He was named Top-40 Early Achievers by The Australian.

Xiaodong Xie received the Ph.D. degree in electrical engineering from the University of Rochester, USA. He was a Senior Scientist with Eastman Kodak Company, NY, USA, from 1994 to 1997, and a Principal Scientist with Broadcom Corporation, CA, USA, from 1997 to 2009. He is currently a Professor with the Department of Computer Science and Technology, Peking University, Beijing, China. His research interests include multimedia SoC design and embedded systems.

Yuan Li (S'12-M'15) received the B.E. degree from South China University of Technology, Guangzhou, China, in 2008, and the Ph. D. degree from Peking University, Beijing, China, in 2015. He currently works with the National Engineering Laboratory for Video Technology at Peking University, China. His research interests include algorithms and VLSI architectures for visual information processing and video coding.

0162-8828 (c) 2021 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.
Authorized licensed use limited to: Peking University. Downloaded on January 12,2022 at 15:26:18 UTC from IEEE Xplore. Restrictions apply.