Devil in the Details: Towards Accurate Single and Multiple Human Parsing

Tao Ruan,[1*] **Ting Liu,**[1*] **Zilong Huang,**[3] **Yunchao Wei,**[†] **Shikui Wei,**[1] **Yao Zhao**[1]
[1]Institute of Information Science, Beijing Jiaotong University, China
[2]University of Illinois at Urbana-Champaign, USA
[3]School of Electronic Information and Communications, Huazhong University of Science and Technology, China

Abstract

Human parsing has received considerable interest due to its wide application potentials. Nevertheless, it is still unclear how to develop an accurate human parsing system in an efficient and elegant way. In this paper, we identify several useful properties, including feature resolution, global context information and edge details, and perform rigorous analyses to reveal how to leverage them to benefit the human parsing task. The advantages of these useful properties finally result in a simple yet effective Context Embedding with Edge Perceiving (CE2P) framework for single human parsing. Our CE2P is end-to-end trainable and can be easily adopted for conducting multiple human parsing. Benefiting the superiority of CE2P, we won the 1st places on all three human parsing tracks in the 2nd Look into Person (LIP) Challenge. Without any bells and whistles, we achieved 56.50% (mIoU), 45.31% (mean AP^r) and 33.34% ($AP^p_{0.5}$) in Track 1, Track 2 and Track 5, which outperform the state-of-the-arts more than 2.06%, 3.81% and 1.87%, respectively. We hope our CE2P will serve as a solid baseline and help ease future research in single/multiple human parsing. Code has been made available at https://github.com/liutinglt/CE2P.

Introduction

Human parsing is a fine-grained semantic segmentation task which aims at identifying the constituent parts (*e.g.* body parts and clothing items) for a human image on pixel-level. Understanding the contents of the human image makes sense in several potential applications including e-commerce, human-machine interaction, image editing and virtual reality to name a few. Currently, human parsing has gained remarkable improvement with the development of fully convolutional neural networks on semantic segmentation.

For the semantic segmentation, researchers have developed many solutions from different views to tackle the challenging dense prediction task. In general, current solutions can be grossly divided into two types: 1) **High-resolution Maintenance** This kind of approaches are attempting to obtain high-resolution features to recover the desired detailed information. Due to consecutive spatial pooling and convolution strides, the resolution of the final feature maps is reduced significantly that the finer image information is lost. To generate high-resolution features, there are two typical solutions, *i.e.* cutting several down-sampling (*e.g.* max pooling) operations (Chen et al. 2016a) and introducing details from low-level feature maps (Chen et al. 2018b). For the latter case, it is usually embedded in an encoder-decoder architecture, in which the high-level semantic information is captured in the encoder and the details and spatial information are recovered in the decoder. 2) **Context Information Embedding** This kind of approaches is devoting to capture rich context information to handle the objects with multiple scales. Feature pyramid is one of the effective ways to mitigate the problem caused by various scales of objects, and atrous spatial pyramid pooling (ASPP) based on dilated convolution (Chen et al. 2018a) and pyramid scene parsing (PSP) (Zhao et al. 2017) are two popular structures. ASPP utilizes parallel dilated convolution layers with different rates to incorporate multi-scale context. PSP designs pyramid pooling operation to integrate the local and global information together for more reliable prediction. Beyond these two typical types, some other works also propose to advantage the segmentation performance by introducing additional information such as edge (Liang et al. 2017; Liu et al. 2017) or more effective learning strategy such as cascaded training (Li et al. 2017b).

Compare with the general semantic segmentation tasks, the challenges of human parsing is to produce finer predictions for every detailed region belonging to a person. Besides, the arms, legs and shoes are further divided into the left side and right side for more precise analysis, which makes the parsing more difficult. Despite the approaches mentioned above showing impressive results in semantic segmentation, it remains unclear how to develop an accurate human parsing system upon these solutions and most previous works did not explore and analyze how to leverage them to unleash the full potential in human parsing. In this work, we target on answering such a question: can we simply formulate a powerful framework for human parsing by exploiting the recent advantages in the semantic segmentation area?

To answer such a question, we conduct a great deal of rigorous experiments to clarify the key factors affecting the

*Equal contribution
†Corresponding author

performance of human parsing. In particular, we perform an analysis of potential gains in mIoU score with different properties. The evaluated useful properties include feature resolution, context information and edge details. Based on the analysis, we present a simple yet effective Context Embedding with Edge Perceiving (CE2P) framework for single human parsing. The proposed CE2P consists of three key modules to learn for parsing in an end-to-end manner: 1) A high-resolution embedding module used to enlarge the feature map for recovering the details; 2) A global context embedding module used for encoding the multi-scale context information; 3) An edge perceiving module used to integrate the characteristic of object contour to refine the boundaries of parsing prediction. Our approach achieves state-of-the-art performance on all three human parsing benchmarks. Those results manifest that the proposed CE2P can provide consistent improvements over various human parsing tasks. The main contributions of our work are as follows:

- We analyze the effectiveness of several properties for human parsing, and reveal how to leverage them to benefit the human parsing task.
- We design a simple yet effective CE2P framework by leveraging the useful properties to conduct human parsing in a simple and efficient way.
- Our CE2P brings a significant performance boost to all three human parsing benchmarks, outperforming the current state-of-the-art method by a large margin.
- Our code is available, which can serve as a solid baseline for the future research in single/multiple human parsing.

Related Work

Human Parsing

The study of human parsing has drawn more and more attention due to the wide range of potential application. The early works (Yamaguchi et al. 2012; Dong et al. 2013; Simo-Serra et al. 2014; Ladicky, Torr, and Zisserman 2013) performed the parsing with CRF framework and utilized the human pose estimation to assist the parsing. A Co-CNN (Liang et al. 2017) architecture was proposed to hierarchically integrate the local and global context information and improved the performance greatly. Recently, SSL (Gong et al. 2017) introduced a self-supervised structure-sensitive loss, which was used for enforcing the consistency between parsing results and the human joint structures, to assist the parsing task. Following previous work, JPPNet (Liang et al. 2018) incorporated the human parsing and pose estimation task into a unified network. With multi-scale feature connections and iterative refinement, the parsing and pose tasks boosted each other simultaneously. Considering the practical application, several current works (Li, Arnab, and Torr 2017; Li et al. 2017a; Zhao et al. 2018) focus on handling the scenario with multiple persons. Usually, it consisted of three sequential steps: object detection (He et al. 2017), object segmentation and part segmentation. Besides, many research efforts (Girshick et al. 2014; Wang et al. 2015; Chen et al. 2016b; Hariharan et al. 2015) have been devoted into the object parts segmentation, which was similar to human parsing. Most of those works leveraged the multi-scale features to enhance the performance.

Semantic Segmentation

Human parsing is a fine-grained semantic segmentation task. Hence, the methods used in human parsing is similar to semantic segmentation. Since the fully convolutional network (FCN) (Long, Shelhamer, and Darrell 2015) has shown numerous improvements in semantic segmentation, many researchers (Chen et al. 2016a; Jégou et al. 2017; Wei et al. 2018; 2017a; 2017b) have made efforts based on the FCN. Several pieces of work (Badrinarayanan, Kendall, and Cipolla 2017; Ronneberger, Fischer, and Brox 2015; Lin et al. 2017), leveraged an encoder-decoder architecture with skip connection to recover the dense feature responses. Another literatures (Chen et al. 2016a; Yu and Koltun 2015; Chen et al. 2018b) exploited the dilated convolution for higher resolution output. Besides, the local and global information are integrated for generating more reliable prediction, such as (Chen et al. 2018a; Zhao et al. 2017).

Context Embedding with Edge Perceiving

In this section, we first provide the architecture of our Context Embedding with Edge Perceiving(CE2P) approach. Within CE2P, we analyze the effectiveness of several key modules motivated from the previous state-of-the-art semantic segmentation models and reveal how they can work together to accomplish the single human parsing task. Then, we give the details of applying the CE2P to address the more challenging multiple human parsing task.

Key Modules of CE2P

Our CE2P integrates the local fine details, global context and semantic edge context into a unified network. The overview of our framework is shown in Fig. 1. Specifically, it consists of three key components to learn for parsing in an end-to-end manner, *i.e.* context embedding module, high-resolution embedding module and edge perceiving module. ResNet-101 is adopted as the feature extraction backbone.

Context Embedding Module Global context information is useful to distinguish the fine-grained categories. For instance, the left and right shoe have relatively high similarity in appearance. To differentiate between the left and right shoe, the global information, like the orientation of leg and body, provides an effective context prior. As we know, feature pyramid is a powerful way to capture the context information. Draw on the previous work PSP (Zhao et al. 2017), we utilize a pyramid pooling module to incorporate the global representations. We perform four average pooling operations on the features extracted from ResNet-101 to generate multiple scales of context features with size 1×1, 2×2, 3×3, 6×6 respectively. Those context features are upsampled to keep the same size with the original feature map by bilinear interpolation, which are further concatenated with the original feature. Then, the 1×1 convolution is employed to reduce the channels and better integrate the multi-scale

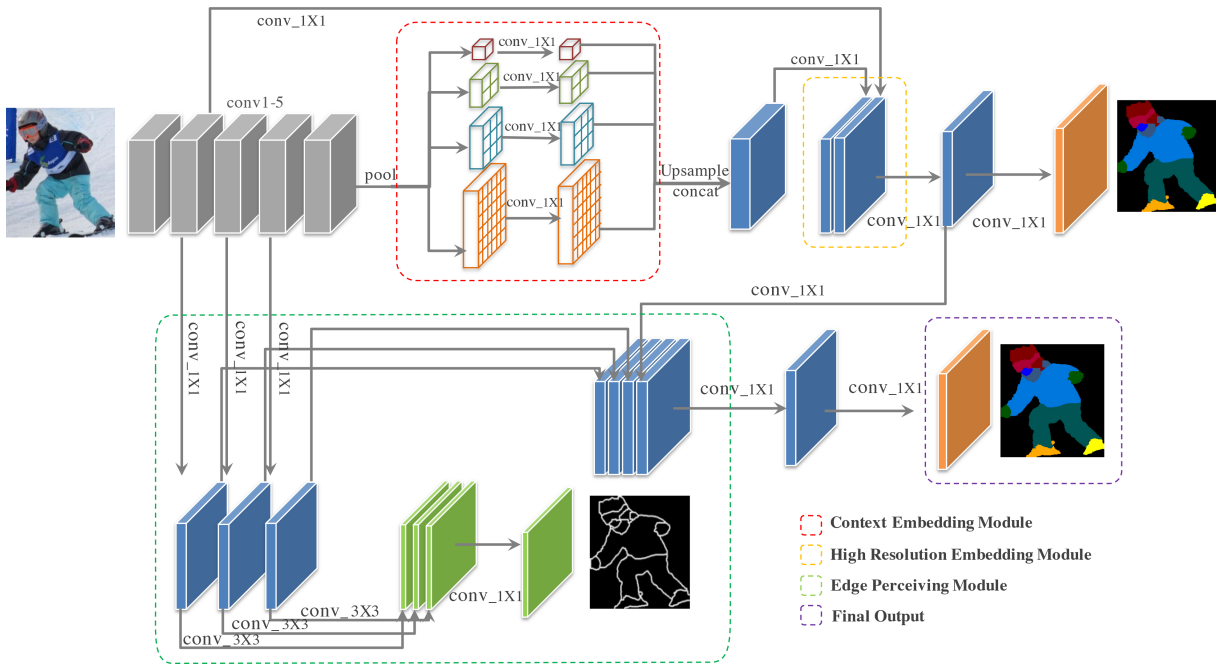

Figure 1: The framework of our proposed CE2P. The overall architecture consists of three key modules: 1) high resolution embedding module 2) global context embedding module; 3) edge perceiving module.

context information. Finally, the output of context embedding module is fed into the following high-resolution module as global context prior.

High-resolution Embedding Module In human parsing, there exist several small objects to be segmented, *e.g.* socks, shoes, sunglasses and glove. Hence, high-resolution feature for final pixel-level classification is essential to generate an accurate prediction. To recover the lost details, we adopt a simple yet effective method which embeds the low-level visual features from intermediate layers as complementary to the high-level semantic features. We exploit the feature from the *conv2* to capture the high-resolution details. The global context feature is upsampled by factor 4 with bilinear interpolation, and concatenated with local feature after channel reduced by 1×1 convolution. Finally, we conduct two sequential 1×1 convolution on the concatenated feature to better fuse the local and global context feature. In this manner, the output of high-resolution module simultaneously acquires high-level semantic and high-resolution spacial information.

Edge Perceiving Module This module aims at learning the representation of contour to further sharp and refine the prediction. We introduce three branches to detect multi-scale semantic edges. As illustrated in Fig. 1, a 1×1 convolution are conducted to *conv2*, *conv3* and *conv4* to generate 2-channel score map for the semantic edge. And then, 1×1 convolution is performed to obtain the fused edge map. Those intermediate features of edge branches, which can capture useful characteristics of object boundaries, are upsampled and concatenated with the features from high-resolution. Finally, 1×1 convolution is performed on the concatenated feature map to predict the pixel-level human parts.

Our CE2P consisting of the three modules is learned with an end-to-end manner. The outputs of CE2P consist of two parsing results and edge prediction. Hence, the loss can be formulated as:

$$L = L_{Parsing} + L_{Edge} + L_{Edge-Parsing}, \qquad (1)$$

where L_{Edge} denotes the weighted cross entropy loss function between the edge map detected by edge module and the binary edge label map; $L_{Parsing}$ denotes the cross entropy loss function between the parsing result from high resolution module and the parsing label; $L_{Edge-Parsing}$ denotes the cross entropy loss function between the final parsing result, which is predicted from the edge perceiving branch, and the parsing label.

Multiple Human Parsing (MHP)

MHP is a more challenging task, which not only needs to classify the semantics of pixels but also identify the instance (*i.e.* one unique person) that these pixels belong to. To achieve high-quality parsing results in the scenario of multiple persons, we design a framework called M-CE2P upon our CE2P and Mask R-CNN (He et al. 2017). As shown in Fig. 2, our M-CE2P leverages three branches, denoted by B_g, B_{l_1}, B_{l_2}, to make predictions from global to local views. The details of the three branches are introduced in the following.

Global Parsing B_g In spite of the CE2P is proposed for single human parsing, we find it shows considerable performance on multiple human images as well. Hence, we first

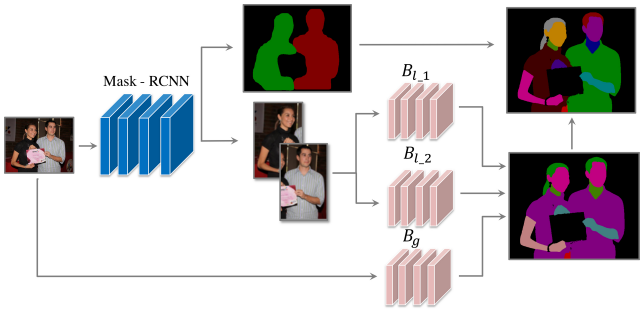

Figure 2: The overall framework of CE2P for Multiple Human Parsing task.

apply it over the entire image for global parsing. For the branch of B_g, we train a CE2P model with the entire images. Then, the output of this branch is leveraged as complementary to the following local parsing. The global parsing branch can provide context information when there exist occlusions among multiple persons. For instance, the same semantic parts form different persons can be easy to tell apart, and the spatial relationship among persons can be captured to handle the circumstance of occlusion. However, it does not concentrate on the relatively small human instances. As a result, body parts belonging to small-scale person are likely ignored by B_g.

Local Parsing with Predicted Instance Masks B_{l_1} To alleviate the problem of global parsing B_g, we consider locating the persons as a preprocessing step to generate accurate parsing results. Towards this end, we propose a two-stage branch devoting to human-level local parsing. Specifically, we employ Mask R-CNN (He et al. 2017) to extract all the person patches in the input image, and resize them to fit the input size of CE2P. And then, all the human-level sub-images are fed into CE2P to training the model for local view. During inference stage, the sub-images with single human instance of input images are extracted by Mask R-CNN, and further fed into the trained model to make parsing predictions. The predicted confidence maps are resized to the original size by bilinear interpolation for the following prediction over the entire image. Finally, the confidence maps for each sub-images are padded with zeros to keep the same size as the confidence map from B_g, and further fused together by element-wise summation on foreground channels and minimization on background channel.

Local Parsing with Ground-truth Instance Masks B_{l_2} Considering that a human instance obtained from the ground-truth instance mask is more approximate to the real single human image, we introduce the branch B_{l_2} to train a model with the data generated from ground-truth instance mask. This branch is rather similar with B_{l_1}, the only difference is that we obtain person patches with the guide of the ground truth bounding boxes in the training stage. With the B_{l_2}, the performance of local parsing can be further boosted. Finally, the predictions generated by the three branches are fused by element-wise summation to obtain the final instance-agnostic parsing result. The predicted instance-agnostic parsing results are further fed into the following process to make the instance-level parsing.

Instance-level Parsing and Label Refinement With the instance-agnostic parsing result obtained from M-CE2P, we consider two aspects to generate the instance-level label, *i.e. instance assignment* for predicting instance-aware results and *label refinement* for solving the shortage of the under-segmentation phenomenon of Mask R-CNN. For instance assignment, we directly apply human masks generated by Mask R-CNN to assign instance-level part label of global body parts. Concretely, parts will be assigned with different part instance labels when they are same category while belonging to different masks. Through the experiments, we find that the parsing mask predicted from our CE2P is more reliable than human instance map. To further validate the reliability of parsing results, we introduce the label refinement by expanding the area of intersection with the neighbor pixels which have same parsing label while exceeding the human instance. For example, some regions of marginal parts(*e.g.* hair, hands) are very likely to be outside of the area of predicted human masks. We use searching based method to alleviate this problem. Specifically, for each border pixel of the part obtained from the assignment step, we use Breadth-First Search to find the pixel that is endowed with an instance-agnostic class label but no part label due to the inaccuracy of segmentation prediction. With the proposed refinement, the body parts excluded by human mask can be effectively included in the final instance-level result.

Experimental Results

Dataset and Metrics

We compare the performance of single human parsing of our proposed approach with other state-of-the-arts on the LIP (Liang et al. 2018) dataset, and we further evaluate the multiple human parsing on CIHP (Gong et al. 2018) and MHP v2.0 (Li et al. 2017a) dataset.

LIP dataset: The LIP (Liang et al. 2017) dataset is used in LIP challenge 2016, which is a large-scale dataset that focuses on single human parsing. There are 50,462 images with fine-grained annotations at pixel-level with 19 semantic human part labels and one background label. Those images are further divided into 30K/10K/10K for training, validation and testing, respectively.

CIHP dataset: CIHP (Gong et al. 2018) provides a dataset with 38,280 diverse human images, in which contains 28,280 training, 5K validation and 5K test images. The images have pixel-wise annotation on 20 categories and instance-level identification.

MHP v2.0 dataset: The MHP v2.0 dataset is designed for multi-human parsing in the wild including 25,403 images with finer categories up to 58 semantic labels. The validation set and test set have 5K images respectively. The rest 15,403 are provided as the training set.

Metrics: We use mean IoU to evaluate the global-level predictions, and use the following three metrics to evaluate the instance-level predictions. $Mean\ AP^r$ computes the area under the precision-recall curve with the limitation of a set of IoU threshold, and figure out the final averaging result,

Table 1: Comparison of CE2P in various module settings on the validation set of LIP. The results are obtained without left-right flipping except for the last row. 'B' means baseline model. 'G', 'H' and 'E' denote global context, high resolution and edge perceiving module, respectively.

Method	bkg	hat	hair	glove	glasses	u-cloth	dress	coat	socks	pants	j-suits	scarf	skirt	face	l-arm	r-arm	l-leg	r-leg	l-shoe	r-shoe	mIoU
B	86.08	63.24	70.14	33.75	29.45	66.15	24.10	51.79	45.70	70.75	21.52	13.56	20.36	73.04	58.24	60.88	49.70	48.06	36.50	36.40	47.97
B+H	86.69	64.65	71.53	33.76	33.30	66.50	25.11	51.52	46.10	71.46	23.62	14.77	21.44	74.08	60.04	62.41	50.36	50.74	36.95	37.19	49.11
B+G	86.61	63.77	71.01	34.90	30.73	67.93	30.89	54.14	46.11	72.21	23.38	13.16	20.16	73.50	59.51	62.12	50.70	50.48	38.63	38.84	49.44
B+E	86.87	63.47	71.56	35.60	31.35	67.06	27.80	52.28	47.85	72.21	23.89	15.60	20.57	74.37	61.04	63.36	53.30	52.70	39.72	40.16	50.04
B+G+H	87.22	65.34	72.13	36.18	31.97	68.86	31.02	55.81	47.35	73.23	26.91	12.28	20.58	74.49	62.95	65.18	56.31	55.59	43.49	43.80	51.54
B+G+H+E (CE2P)	87.41	64.62	72.07	38.36	32.20	68.92	32.15	55.61	48.75	73.54	27.24	13.84	22.69	74.91	64.00	65.87	59.66	58.02	45.70	45.63	52.56
CE2P (Flipping)	**87.67**	65.29	**72.54**	**39.09**	**32.73**	**69.46**	**32.52**	**56.28**	**49.67**	**74.11**	27.23	**14.19**	22.51	**75.50**	**65.14**	**66.59**	**60.10**	**58.59**	**46.63**	**46.12**	**53.10**

which is first introduced in (Hariharan et al. 2014); AP^p (Li et al. 2017a) computes the pixel-level IoU of semantic part categories within a person, instead of global circumstance; PCP (Li et al. 2017a) elaborates how many body parts are correctly predicted of a certain person, guided with pixel-level IoU.

Implement Details

We implement the proposed framework in PyTorch (Paszke et al. 2017) based on (Huang et al. 2018), and adopt ResNet-101 (He et al. 2016) as the backbone network. The input size of the image is 473×473 during training and testing. We adopt the similar training strategies with Deeplab (Chen et al. 2018b), *i.e.* "Poly" learning rate policy with base learning rate 0.007. We fine-tune the networks for approximately 150 epochs. For data augmentation, we apply the random scaling (from 0.5 to 1.5), cropping and left-right flipping during training. Note that the edge annotation used in the edge perceiving module is directly generated from the parsing annotation by extracting border between different semantics.

Table 2: Comparison of performance on the validation set of LIP with state-of-arts methods.

Method	pixel acc.	mean acc.	mIoU
DeepLab (VGG-16)	82.66	51.64	41.64
Attention	83.43	54.39	42.92
DeepLab (ResNet-101)	84.09	55.62	44.80
JPPNet (Liang et al. 2018)	86.39	62.32	51.37
CE2P	**87.37**	**63.20**	**53.10**

Single Human Parsing

To investigate the effectiveness of each module, we report the performance under several variants of CE2P in Tab. 1. We begin the experiment with a baseline model without any proposed modules. For our baseline, the prediction is directly performed on the final feature map extracted from ResNet-101. The resolution of the final feature map is $1/16$ to the input size. The results are shown in Tab. 1, and we can see the baseline model reaches 47.97% accuracy. Some failure examples are shown in Fig. 3. Observing the per-classes performance and the visualized results, there exist the following problems. 1) Big-size objects have the discontinuous prediction. For instance, the dress is always parsed as a upper-clothes and skirt, and the jumpsuit is separated into a upper-clothes and pants. 2) The confusable categories are hard to distinguish. *i.e.* the coat and upper-clothes. 3) The left and right parts are easily confused, which frequently occurs in the back-view human body and the front-view body with legs crossed.

Global Context Embedding Module To evaluate the effectiveness of each module, we first conduct experiments by introducing a global context embedding module. In our architecture, we leverage the pyramid pooling (Zhao et al. 2017) to generate multi-scale context information. As shown in Tab. 1, we can find it brings about 1.5% improvements on mIoU, which demonstrates that the multi-scale context information can assist the fine-grained parsing. Particularly, it shows significant boosts (nearly 7%) in the class of dress. Since the long-range context information can provide the more discriminated characteristic, the global context embedding module is helpful for the big-size objects.

High Resolution Embedding Module To figure out the importance of the high resolution, we conduct experiments by further introducing a high resolution module. From Tab. 1, we can find that the performance gains nearly 2% improvement with the high resolution embedding module. As human parsing is a fine-grained pixel-wise classification, it requires a lot of detailed information to identify the various small-size parts accurately. With high resolution embedding module, the features from shallow and high-resolution layers provide more details not available in deep layers. The results demonstrate the effectiveness as well.

Edge Perceiving Module Finally, we report the performance with the edge perceiving module in Tab. 1. Based on the above two modules, appending edge perceiving module still brings nearly 1% boosts. That's the contours of the parts can be underlying constraints during separating the semantic parts from a human body. In addition, the features from multiple edge branches carrying various details of the objects can further promote the human parts prediction. Finally, fusing with the flipped images gives 0.6% gain.

Comparison with State-of-the-Arts We evaluate the performance of CE2P on the validation dataset of LIP and compare it to other state-of-the-art approaches. The results are reported in Tab. 2. First, we note that the mIoU of our CE2P significantly outperforms other approaches. The improvement over the state-of-the-art method validates the effectiveness of our CE2P for human parsing. When compar-

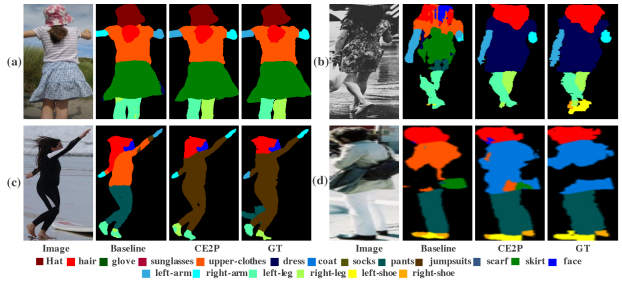

Figure 3: Some issues of baseline model on LIP dataset. (a) The left arm (leg) is wrongly predicted as right arm (leg). (b) The dress is separated into a upper-clothes and skirt. (c) The jumpsuits is parsed into a upper-clothes and pants. (d) The coat is mislabeled as upper-clothes.

Figure 4: A visualized example for Multiple Human Parsing task. The areas surrounded by white bounding boxes indicate the failure cases of predicted results.

ing with the current state-of-the-art approach JPPNet, our method exceeds by 1.73% in terms of mIoU. In particular, the performance on the small-size categories, *i.e.* 'socks' and 'sunglasses', yields obviously improvement. Thanks to the high resolution embedding and edge perceiving module, the details and characteristic of small objects can be captured for further pixel-level classification. Besides, the JPPNet achieves the performance of 51.37% by utilizing extra pose annotation with a complex network structure. Nevertheless, our CE2P obtains better performance with a simpler network structure and no need for extra annotation.

Table 3: Comparison of the performance on the test set of single and multiple human parsing datasets. Here the subscript m in X_m means the mean value of X.

Method	pixel acc.	mean acc.	mIoU
Single human parsing (Track 1)			
ours	**88.92**	**67.78**	**57.90**
ours (single model)	88.24	67.29	56.50
JD_BUPT	87.42	65.86	54.44
AttEdgeNet	87.40	67.17	54.17
Method	mIoU	AP^r_m	$AP^r_{0.5}$
Multiple human parsing on CIHP (Track 2)			
ours	**63.77**	**45.31**	**50.94**
DMNet(2nd place)	61.51	41.50	46.12
PGN (Gong et al. 2018)	55.80	33.60	35.80
Method	$PCP_{0.5}$	$AP^p_{0.5}$	AP^p_m
Multiple human parsing on MHP v2.0 (Track 5)			
ours	**41.82**	**33.34**	**42.25**
S-LAB(2nd place)	38.27	31.47	40.71
NAN (Zhao et al. 2018)	32.25	25.14	41.78

Multiple Human Parsing

We provide abundant experimental results on two large datasets of multiple human parsing task, named CIHP and MHP v2.0. The results are exhibited in Tab. 4. Due to space limitation, we only show detailed results on the more challenging MHP v2.0 dataset. The detailed explanation will be given in the following paragraphs.

Label Refinement A remarkable improvement attributes the success to the label refinement operation. As Tab. 4 shows, it exactly brings performance boosting, despite of the combination strategies. As we mentioned before, the results provided by Mask R-CNN are likely to ignore some partial area of marginal body parts, especially on complex images. However, CE2P may catch these parts from the localized human sub images. Therefore, the refinement operation can alleviate the under segmentation problem. For clarity, the following analyses are all based on refined results.

Comparison with Various Combination Strategies To prove the effectiveness of the multi-branch fusion strategies, we perform experiments with various branch combinations. From Tab. 4, we can notice that the results produced by the double-branch model are better than that produced by the single-branch model, and the all-branch model, *i.e.* M-CE2P, catch the best performance on most of the metrics. Especially on more convincing human-centric metrics like AP^p and PCP, M-CE2P shows out a significant performance boosting than all the other models. It proves that branches can make complements with each other. As mentioned in the previous section, the branch B_g trained with the entire image lacks the ability to grab small-scale persons in a scene, such as the first image in Fig. 4. Hence, it only achieves the performance of 24.04% in terms of $AP^p_{0.5}$. Nevertheless, this shortcoming can be compensated by B_{l_1} and B_{l_2} to capture a more precise view of local context. On the other hand, the benefits from B_g are still cannot be ignored. The global context that B_g has makes a performance improvement of 2.08% in terms of $AP^p_{0.5}$ than the result only utilizes B_{l_1} and B_{l_2}. Finally, the all-branch fused M-CE2P can reach the best performance under the AP^p and PCP metrics.

Comparison with State-of-the-Arts Compared to other state-of-the-art methods, our M-CE2P model still maintains a large-margin leading edge on major metrics, as Tab. 3 and Tab. 4 illustrate. On validation set of MHP v2.0, our proposed model outperforms than (Zhao et al. 2018) 9.64%, 9.40% in terms of $AP^p_{0.5}$ and $PCP_{0.5}$, respectively; Furthermore, on the test set of MHP v2.0, we outperform than (Zhao

Table 4: Comparisons on MHP v2.0 validation dataset

Method	mIoU	$AP^r_{0.5}$	Mean AP^r	$AP^p_{0.5}$	Mean AP^p	$PCP_{0.5}$	Mean PCP
NAN (Zhao et al. 2018)	-	-	-	24.83	42.77	34.37	-
B_g	38.25	27.32	24.96	21.63	37.13	32.59	34.94
B_{l_1}	40.18	30.82	27.95	24.98	38.72	36.42	37.54
B_{l_2}	40.30	30.75	28.04	24.46	38.60	36.09	37.46
$B_{l\_1} + B_g$	40.48	30.63	27.79	29.62	40.52	39.29	38.46
$B_{l\_2} + B_g$	40.52	30.67	27.89	29.29	40.56	39.13	38.50
$B_{l\_1} + B_{l\_2}$	41.05	31.63	28.82	28.71	40.37	39.32	38.90
$B_{l\_1} + B_{l\_2} + B_g$(M-CE2P)	41.11	31.50	28.60	30.92	41.29	40.58	39.32
B_g with refinement	38.25	29.70	26.94	24.04	38.21	35.09	36.36
B_{l_1} with refinement	40.18	33.69	30.34	28.20	40.03	39.56	39.21
B_{l_2} with refinement	40.30	33.66	30.41	27.43	39.84	39.09	39.11
$B_{l\_1} + B_g$ with refinement	40.48	33.39	30.08	32.84	41.90	42.32	40.19
$B_{l\_2} + B_g$ with refinement	40.52	33.54	30.22	32.62	41.88	42.14	40.19
$B_{l\_1} + B_{l\_2}$ with refinement	41.05	**34.58**	**31.24**	32.39	41.75	42.64	40.66
$B_{l\_1} + B_{l\_2} + B_g$(M-CE2P) with refinement	**41.11**	34.40	30.97	**34.47**	**42.70**	**43.77**	**41.06**

et al. 2018) 8.20%, 9.57% and 0.47% in terms of $AP^p_{0.5}$, $PCP_{0.5}$ and mean AP^p, respectively; On the test set of CIHP, our M-CE2P performs 63.77%, 45.31%, 50.94% in terms of mIoU, mean AP^r and $AP^r_{0.5}$, respectively, which all outperform than (Gong et al. 2018). Above all, we investigate and combine several useful properties to improve the parsing results. Each module plays an important role and makes improvements for the final results. Benefiting from our parsing network, the prediction on the whole multi-person image already has a comparable result with baseline methods. Based on the accurate parsing results, our fusion strategy and label refinement can make further boost.

Visualization Some visualized results with their failure cases are shown in Fig. 4. Generally speaking, our M-CE2P can handle a rather complex scene with multiple persons, and produce a satisfactory result. However, there are also some failure cases produced by our M-CE2P. From the perspective of fusion strategy, it has following problems: 1) Under some circumstances, the B_g brings too much negative confidence to parts belong to human far away from camera, so that the confidence provided by local parsing may be drastically reduced; 2) When acting sub-image fusion, the local information of tightly closed humans may be disturbed with each other. For example, in bottom line of Fig. 4, the man with black suit make the edge of woman's skirt strongly be mistaken as other semantics.

Results in Challenge

With our CE2P framework, we achieved the 1st places of all three human parsing tracks in the 2nd Look into Person (LIP) Challenge. Tab. 3 shows a few of the results on the single human parsing lead-board. By integrating the results from three models with different input size, our CE2P achieved 57.9%. More importantly, our single model already attained the state-of-arts performance without any bells and whistles. Besides, we design a M-CE2P upon our CE2P for multiple human parsing with three branches to predict from global to local view. Benefiting from the M-CE2P model, we achieved high performance in all the multiple human parsing benchmarks without refinement process, *i.e.* CIHP and MHP v2.0, respectively. Specifically, we yielded 45.31% of Mean AP^r and 33.34% of $AP^p_{0.5}$, which outperform the second place more than 3.81% 1.87%, respectively.

Conclusion

In this paper, we investigate several useful properties for human parsing, including feature resolution, global context information and edge details. We design a simple yet effective CE2P system, which consists of three key modules to incorporate the context and detailed information into a unified framework. Specifically, we use a high-resolution embedding module to capture details from shallow layer, a global context embedding module for multi-scale context information, and an edge perceiving module to constrain object contours. For multiple human parsing, we fuse three CE2P based model to generate global parsing prediction, and use a straight-forward way to produce the instance-level result with the guide of human mask. The experimental results demonstrate the superiority of the proposed CE2P.

Acknowledgment

This work was supported in part by National Key Research and Development of China (No.2017YFC1703503), National Natural Science Foundation of China (No.61532005, No.61572065), Fundamental Research Funds for the Central Universities (No. 2017YJS048, No. 2018JBZ001), Joint Fund of Ministry of Education of China and China Mobile (No.MCM20160102), IBM-ILLINOIS Center for Cognitive Computing Systems Research (C3SR) - a research collaboration as part of the IBM AI Horizons Network.

References

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

Chen, L.; Barron, J. T.; Papandreou, G.; Murphy, K.; and Yuille, A. L. 2016a. Semantic image segmentation with

task-specific edge detection using cnns and a discriminatively trained domain transform. In *CVPR*, 4545–4554.

Chen, L.; Yang, Y.; Wang, J.; Xu, W.; and Yuille, A. L. 2016b. Attention to scale: Scale-aware semantic image segmentation. In *CVPR*, 3640–3649.

Chen, L.; Papandreou, G.; Kokkinos, I.; Murphy, K.; and Yuille, A. L. 2018a. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *TPAMI* 40(4):834–848.

Chen, L.; Zhu, Y.; Papandreou, G.; Schroff, F.; and Adam, H. 2018b. Encoder-decoder with atrous separable convolution for semantic image segmentation. *arXiv:1802.02611*.

Dong, J.; Chen, Q.; Xia, W.; Huang, Z.; and Yan, S. 2013. A deformable mixture parsing model with parselets. In *ICCV*, 3408–3415.

Girshick, R. B.; Donahue, J.; Darrell, T.; and Malik, J. 2014. Rich feature hierarchies for accurate object detection and semantic segmentation. In *CVPR*, 580–587.

Gong, K.; Liang, X.; Zhang, D.; Shen, X.; and Lin, L. 2017. Look into person: Self-supervised structure-sensitive learning and a new benchmark for human parsing. In *CVPR*, 6757–6765.

Gong, K.; Liang, X.; Li, Y.; Chen, Y.; Yang, M.; and Lin, L. 2018. Instance-level human parsing via part grouping network. *arXiv:1808.00157*.

Hariharan, B.; Arbeláez, P.; Girshick, R.; and Malik, J. 2014. Simultaneous detection and segmentation. In *ECCV*, 297–312.

Hariharan, B.; Arbeláez, P.; Girshick, R.; and Malik, J. 2015. Hypercolumns for object segmentation and fine-grained localization. In *CVPR*, 447–456.

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

He, K.; Gkioxari, G.; Dollár, P.; and Girshick, R. 2017. Mask r-cnn. In *ICCV*, 2980–2988.

Huang, Z.; Wei, Y.; Wang, X.; and Liu, W. 2018. A pytorch semantic segmentation toolbox.

Jégou, S.; Drozdzal, M.; Vázquez, D.; Romero, A.; and Bengio, Y. 2017. The one hundred layers tiramisu: Fully convolutional densenets for semantic segmentation. In *CVPR*.

Ladicky, L.; Torr, P. H. S.; and Zisserman, A. 2013. Human pose estimation using a joint pixel-wise and part-wise formulation. In *CVPR*, 3578–3585.

Li, Q.; Arnab, A.; and Torr, P. H. S. 2017. Holistic, instance-level human parsing. *arXiv:1709.03612*.

Li, J.; Zhao, J.; Wei, Y.; Lang, C.; Li, Y.; and Feng, J. 2017a. Towards real world human parsing: Multiple-human parsing in the wild. *arXiv:1705.07206*.

Li, X.; Liu, Z.; Luo, P.; Change Loy, C.; and Tang, X. 2017b. Not all pixels are equal: Difficulty-aware semantic segmentation via deep layer cascade. In *CVPR*, 3193–3202.

Liang, X.; Xu, C.; Shen, X.; Yang, J.; Tang, J.; Lin, L.; and Yan, S. 2017. Human parsing with contextualized convolutional neural network. *TPAMI* 39(1):115–127.

Liang, X.; Gong, K.; Shen, X.; and Lin, L. 2018. Look into person: Joint body parsing & pose estimation network and A new benchmark. *TPAMI*.

Lin, G.; Milan, A.; Shen, C.; and Reid, I. D. 2017. Refinenet: Multi-path refinement networks for high-resolution semantic segmentation. In *CVPR*, 5168–5177.

Liu, T.; Wei, Y.; Zhao, Y.; Liu, S.; and Wei, S. 2017. Magicwall: Visualizing room decoration. In *ACM Multimedia*, 429–437.

Long, J.; Shelhamer, E.; and Darrell, T. 2015. Fully convolutional networks for semantic segmentation. In *CVPR*, 3431–3440.

Paszke, A.; Gross, S.; Chintala, S.; Chanan, G.; Yang, E.; DeVito, Z.; Lin, Z.; Desmaison, A.; Antiga, L.; and Lerer, A. 2017. Automatic differentiation in pytorch. In *NIPS-W*.

Ronneberger, O.; Fischer, P.; and Brox, T. 2015. U-net: Convolutional networks for biomedical image segmentation. In *MICCAI*, 234–241.

Simo-Serra, E.; Fidler, S.; Moreno-Noguer, F.; and Urtasun, R. 2014. A high performance CRF model for clothes parsing. In *ACCV*, 64–81.

Wang, P.; Shen, X.; Lin, Z. L.; Cohen, S.; Price, B. L.; and Yuille, A. L. 2015. Joint object and part segmentation using deep learned potentials. In *ICCV*, 1573–1581.

Wei, Y.; Feng, J.; Liang, X.; Cheng, M.-M.; Zhao, Y.; and Yan, S. 2017a. Object region mining with adversarial erasing: A simple classification to semantic segmentation approach. In *CVPR*.

Wei, Y.; Liang, X.; Chen, Y.; Shen, X.; Cheng, M.-M.; Feng, J.; Zhao, Y.; and Yan, S. 2017b. Stc: A simple to complex framework for weakly-supervised semantic segmentation. *TPAMI* 39(11):2314–2320.

Wei, Y.; Xiao, H.; Shi, H.; Jie, Z.; Feng, J.; and Huang, T. S. 2018. Revisiting dilated convolution: A simple approach for weakly-and semi-supervised semantic segmentation. In *CVPR*, 7268–7277.

Yamaguchi, K.; Kiapour, M. H.; Ortiz, L. E.; and Berg, T. L. 2012. Parsing clothing in fashion photographs. In *CVPR*, 3570–3577.

Yu, F., and Koltun, V. 2015. Multi-scale context aggregation by dilated convolutions. *arXiv:1511.07122*.

Zhao, H.; Shi, J.; Qi, X.; Wang, X.; and Jia, J. 2017. Pyramid scene parsing network. In *CVPR*, 6230–6239.

Zhao, J.; Li, J.; Cheng, Y.; Zhou, L.; Sim, T.; Yan, S.; and Feng, J. 2018. Understanding humans in crowded scenes: Deep nested adversarial learning and A new benchmark for multi-human parsing. *arXiv:1804.03287*.