BiEPNet: Bilateral Edge-perceiving Network for High-Resolution Human Parsing

Qiqi Gong[*] Yao Zhao[†] Yunchao Wei[‡]

Institute of Information Science, Beijing Jiaotong University
Beijing Key Laboratory of Advanced Information Science and Network

Abstract

Human parsing is a fundamental task aimed at segmenting human images into distinct body parts and holds vast potential applications. Nowadays, the advancement of image-capturing devices has led to a growing number of high-resolution human images. Receptive field, details loss and memory usage are a triplet of contradictions in high-resolution scenarios. Existing human parsing methods designed for low-resolution inputs struggle to process high-resolution images efficiently due to their massive demands for computation and memory. Some methods save resources by overwhelmingly downsampling or encoding high-resolution inputs at the cost of poor performance on details. To resolve the issues above, we propose the Bilateral Edge-Perceiving Network (BiEPNet), consisting of a resources-friendly semantic-perceiving branch to acquire sufficient global information and a simple yet effective edge-perceiving branch used to refine details. The attention mechanism is utilized to simultaneously enhance the perception of context and details, leading to better performance on the boundary regions. To verify the effectiveness of BiEPNet, we contribute a high-resolution human parsing dataset, Human4K, containing 4,000 images with more than five million pixels. Extensive experiments on Human4K demonstrate that our method outperforms state-of-the-art methods while maintaining memory efficiency.

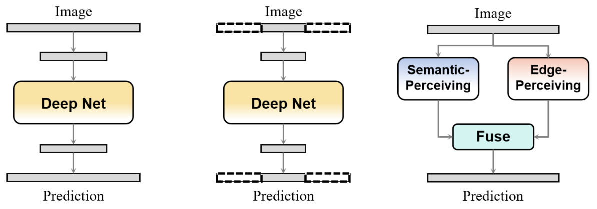
Index Terms: Human-centered computing—Human Parsing—High-Resolution

1 Introduction

Human parsing, which is a fine-grained semantic segmentation task, aims to assign pixel-level labels for an input human image [27]. Serving as the core of human understanding tasks, human parsing has drawn great interest from researchers in the deep learning era and has been widely used in many sectors, *e.g.* fashion, electronic commerce, safety, image editing, etc.

Although acquired great success, existing human parsing methods are universally trained and validated on human images with pixels around 100 thousand pixels. With the advancement of image-capturing devices, high-resolution human images have largely emerged on social networks. Most cameras and displays have a resolution that surpasses 1080P (1920 × 1080, 2 million pixels), while those of a commercial nature have been upgraded to 4K UHD (3840 × 2160, 8 million pixels) [8]. Under this situation, existing human parsing methods can hardly be generalized to high-resolution scenarios due to memory and computational limitations.

DeepLabv3+ [4] is one of the impactful semantic segmentation models that have been widely used in some specific tasks [14,25,29]. Based on the atrous spatial pyramid pooling (ASPP) module proposed in [2], DeepLabv3+ introduced an encoder-decoder structure where the low-level features extracted by the backbone are involved in the decoding process. Based on empirical studies, we find that this kind of skip-layer design could help the model discriminate the foreground from the background well, but details within the subject could be ignored. Additionally, a sharp increase in the usage of GPU memory brought by shortcuts [28] is not tolerated in high-resolution scenarios.



(a) Downsample (b) Sliding Windows (c) Ours

Figure 1: A comparison of potential high-resolution human parsing solutions and ours. The downsample-upsample process in Fig. 1(a) brings details loss; Sliding windows in Fig. 1(b) break the context information; Our framework is composed of a semantic-perceiving branch and an edge-perceiving branch to provide context information and keep attention to details. Information from the two branches is fused to commonly contribute to the final result.

Downsampling and sliding windows are two intuitive solutions to dealing with the dilemma. As shown in Fig. 1, downsampling-based methods first downsample high-resolution human images into low-resolution ones, and the low-resolution predicting results are directly upsampled to obtain high-resolution segmentation maps. Sliding-window-based methods crop the entire image to several parts and make predictions for them respectively. However, these two methods have some drawbacks. Simply downsampling high-resolution images brings loss of details, especially for small-scale objects, and finally results in performance dropping. Sliding windows feed the network with a patch of the whole image, saving memory usage at the cost of the consistency of the context information.

Recently, some resource-saving segmentation models have been proposed, inspiring the research on processing high-resolution images, among which structures with dual sub-networks are plain but competitive. BiSeNet [34] is initially designed to realize real-time semantic segmentation; but due to its excellent trade-off between memory usage and performance, it is ideal to utilize it as a baseline for high-resolution semantic segmentation. Encouraged by the design of bilateral paths, ISDNet [13] suggests combining the deep segmentation net (where high-resolution images are downsampled) with a shallow net and gradually fuses the features to generate the final high-resolution outputs. Similarly, PGNet [32] introduces a pyramid grafting network to fuse features from the deep convolutional network and transformer architecture for high-resolution salient ob-

[*]e-mail: gongqiqi@bjtu.edu.cn
[†]e-mail: yzhao@bjtu.edu.cn
[‡]e-mail: yunchao.wei@bjtu.edu.cn

ject detection. We argue that these methods are sub-optimal. The context path in BiSeNet overwhelmingly encodes the input into 1/32 of the initial sizes, failing to maintain details. Meanwhile, the 'DeepNet+ShallowNet' structures in ISDNet and PGNet could confuse the network during the fusion of features due to the independent semantics of the two branches. Also, simple downsampling with the aim of computing reduction for the deep net rehearses details loss.

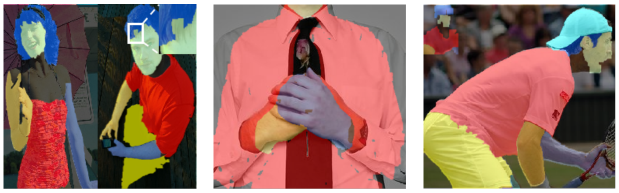
To tackle the problems mentioned above, we propose a novel **Bi**lateral **E**dge-**P**erceiving **Net**work (BiEPNet, abstracted in Fig. 1(c)) oriented towards high-resolution human parsing in an end-to-end training and validating manner. Our proposed BiEPNet is composed of a memory-saving semantic-perceiving branch to provide abundant global context information and an efficient edge-perceiving branch to ensure the focus on details. Differing from typical bilateral structures, where high-resolution input images are often heavily downsampled or encoded, we have successfully preserved resolutions for both input images and feature maps, striking a balance among receptive fields, details, and memory usage. Moreover, our edge-perceiving branch effectively learns the representation of high-frequency regions, thereby augmenting attention to details and yielding significant improvements in overall high-resolution human parsing. Compared to some generic semantic segmentation methods, we consume less GPU memory but obtain better performance. In order to verify the effectiveness of our BiEPNet, we also contribute a high-resolution and well-annotated human parsing dataset Human4K and conduct comprehensive experiments on it. Contributions of this paper can be summarized as below:

- We propose a resource-saving yet effective framework **Bi**lateral **E**dge-**P**erceiving **Net**work (BiEPNet) to directly process high-resolution human images, while still maintaining either the receptive field or attention to details.
- We contribute a high-resolution human parsing dataset Human4K, including 4,000 human images with over 5 million pixels and accurate annotations.
- Extensive experiments demonstrate that our method achieves superior performance compared to existed methods while consuming fewer memory resources during inference.

2 Related Works

2.1 Semantic Segmentation

Semantic segmentation is a significant task in computer vision that has achieved remarkable success. The use of deep convolutional neural networks has enabled the development of FCN-based [24] semantic segmentation models, which have been tested on various benchmarks and have yielded impressive results [4,26,38].DeepLabv3+ [4] utilized an atrous spatial pyramid pooling module with an encoder-decoder structure to acquire sufficient information from multiple levels. PSPNet [38] introduced a pyramid pooling module to capture multi-scale context. Although having obtained extraordinary performances, these methods require massively spatial and temporal computation costs due to their structures. A linear increase in computation costs with the input size results in these architectures struggling to effectively process high-resolution images. To address the problems raised by the heavyweight models, some lightweight but competitive segmentation models were proposed. BiSeNetV1 [34] introduced a bilateral structure composed of a context path and a spatial for high-level information and low-level details respectively. Based on BiSeNet, STDC [11] presents a lightweight backbone to achieve faster speed and higher accuracy. However, the aforementioned methods do not robustly handle high-resolution segmentation well. Differently, we propose a novel high-resolution human parsing framework that incorporates a semantic-perceiving branch and an edge-perceiving branch, yielding competitive or even superior results compared to both lightweight and heavyweight models.



(a) Noise (b) Incomplete (c) Mismatch

Figure 2: Examples of some drawbacks of existing human parsing dataset. (a) Annotations in available datasets are noisy, and some key parts are not annotated; (b) Some images display an incomplete human body, disturbing the learning process; (c) A mismatch between images and annotations characterized by the number of objects is different from that of annotations.

2.2 Human Parsing

Human parsing has experienced significant development in the era of deep learning. Context learning has emerged as one of the prominent paradigms employed for conducting human parsing. Chen et al. [3] exploited the attention mechanism in modeling human structures and by then Chen et al. [6] reculated regional feature weights for segmentation results. Wang et al. [30] designed a parallel-connected network to aggregate features of different resolutions to acquire stronger context information. Other studies tried to use auxiliary information as additional supervision to refine human parsing results. Ruan et al. [27] proposed CE2P, leveraging global context information and edge details to benefit human parsing. Due to its superb performance, CE2P has been generally used as the baseline for subsequent works. Zhang et al. [36, 37] introduced pose information as a piece of auxiliary information to further help the model understand the relationships among human parts. SCHP [20] utilized self-correction methods to reduce the noise for annotations in LIP [12], achieving the 1st place in the CVPR2019 LIP Challenge. CDGNet [23] suggested that human parts are highly related to their positions, generating horizontal and vertical class distribution labels as supervision signals.

2.3 High-resolution Image Processing

While downstream tasks in computer vision have been extensively researched, the majority of focus has been on low-resolution input and output, leaving high-resolution scenarios insufficiently explored. CascadePSP [8] is one of the earliest work studied high-resolution image processing. It is a model-agnostic structure that continuously refines the output from the previous stage. Besides, some work [19, 35, 39] tried to refine the boundary performance to promote overall accuracy. Recently, PGNet [32] introduced a framework fusing features from both CNN and Transformer to leverage their advantages on high-resolution salient object detection. Although these works have shown promising results in binary classification tasks, the dynamics change when dealing with multi-class classification, primarily due to the presence of inter-class conflicts. The global-local framework is a popular paradigm for solving high-resolution segmentation. GLNet [5] combines the context information from the global branch and details from the local branch to improve the segmentation results. Similarly, a classification branch was proposed in PPN [31] to select important patches and fuse with global images. MagNet [17] refined rough results progressively with multi-scale. ISDNet [13] proposed a 'DeepNet+ShallowNet' structure to learn information from different scales. However, the above solutions involve redundant calculation and slow prediction speed, and the information interaction between branches could confuse the network. In contrast, we endow our network with independent abilities contributing to the segmentation results in a resource-saving way.

3 High-Resolution Human Parsing Dataset

Existed human parsing datasets. Current publicly available human parsing datasets are relatively scarce compared to generic semantic segmentation datasets. ATR [21] and LIP [12] are two standard datasets commonly used in human parsing research. The LIP dataset comprises a total of 50,462 images, divided into 30,462/10,000/10,000 for training/validating/testing and labeled with 20 semantic classes. In contrast, the ATR dataset consists of 17,700 images distributed across 18 categories. These datasets are characterized by their large scale, but they typically exhibit low resolution, with most images containing hundreds of thousands of pixels at most. Moreover, they have some drawbacks impeding the learning of representations (see Fig. 2). First, labels are usually noisy. Second, humans displayed in some images are fragmented. Third, there exists a mismatch between images and annotations where the number of objects is different from that of annotations.

High-Resolution Human Parsing Dataset. To bridge the gap between high-resolution scenarios and human parsing, we contribute a high-resolution human parsing dataset Human4K. This dataset was initially collected by Maadaa Data, including over 10,000 images. We select 4,000 images to form the High-Resolution Human Parsing dataset. Each image contains at least 5 million pixels, approaching 4K resolution, and some images have over 20 million pixels, reaching the standards of contemporary commercial cameras. We follow the categories in LIP for the first 20 categories and add two categories named *Torso-skin* and *Else*(initially labeled as necklace, bracelet, etc) and thus 22 categories in total. Images are randomly split into 2,500/500/1000 training/validating/testing.

Compared to LIP where person instances are cropped from Microsoft COCO [22], all images in Human4K are captured in real scenes and the complexities of scenes are different. Each sample in Human4K strictly displays just one person, avoiding the mismatch between human instances and annotations. To make our network more human-centric, nearly all images include *Face* category, but diverse in appearances, postures and viewpoints. Our Human4K addresses the aforementioned drawbacks of existing datsets.

4 Methods

4.1 Architecture Overview

Based on our Human4K, we propose a Bilateral Edge-Perceiving Network, BiEPNet. As shown in Fig. 3, our BiEPNet comprises a semantic-perceiving branch and an edge-perceiving branch. The two branches are designed with distinct objectives, collaborating closely to contribute to the parsing result. The semantic-perceiving branch aims to provide sufficient semantic and context information with a segmentation network (Sect. 4.2), and the edge-perceiving branch enhances the attention to the boundary parts (Sect. 4.3). Information from the two branches is fused in an attention-based method to form the final result. Unlike ISDNet [13], we directly feed both branches with high-resolution input images without downsampling operation.

4.2 Semantic-perceiving Branch

Our semantic-perceiving branch serves as a strong global and context information preceptor, following the design of DeepLabv3 [2]. Given the image $I \in R^{3\times H\times W}$, where H and W are the height and width of the image respectively, the backbone will output a feature map from the last block denoted as $f \in R^{C\times h\times w}$ where $(h,w) = \left(\frac{H}{8}, \frac{W}{8}\right)$. The feature f is then decoded with an ASPP head [4] to obtain strong semantic information $f_{sp} \in R^{C_{sp}\times h\times w}$. An FCN [24] head can be added on the top of the semantic-perceiving branch to acquire the coarse segmentation map $\mathbf{S_{deep}} \in R^{cls\times h\times w}$, where cls is the number of categories.

Compared with the context branch in BiSeNet [34] and the deep net in ISDNet [13], our semantic-perceiving branch captures abundant semantic information without compromising the requirement for dense prediction. BiSeNet overwhelmingly encodes the input image with output stride 32 (ours is 8), and ISDNet directly downsamples the input image into $I' \in R^{3\times\frac{H}{4}\times\frac{W}{4}}$ to preserve memory, degrading details in the high-resolution input images.

4.3 Edge-perceiving Branch

The quality of prediction results on boundary parts is one of the distinctive challenges in human parsing [23, 27, 36, 37]. Compared to complex semantic information, edge information is a relatively low-level type of information. Some operators [1, 18] can obtain coarse edge maps according to input RGB images. In the deep learning era, these operators are simulated with unlearnable convolutional operators [8, 11]. We argue that edge information could be well perceived with a few convolutional layers, and leveraging this information can significantly enhance overall parsing results.

Based on empirical studies, we notice that low-level features in DeepLabv3+ [4] struggle to recognize details within the effectively discriminated foreground and background(see Fig. 4(a)). In the spatial path of BiSeNet [34], which consists of several convolutional layers, edge parts are perceived but they make weak contributions to the consequent process (see Fig. 4(b)). To enhance the utilization of boundary information for refining parsing results, we employ several convolution layers as the boundary extractor and reinforce the features with the self-attention mechanism, thereby constructing our edge-perceiving branch. Denote features extracted by consequent convolution layers as $f_s \in R^{C_s\times h\times w}$. In order to relate detailed edge information with other long-range and short-range positions spatially, we utilize the interlaced sparse self-attention mechanism [16] on f_s and finally acquired features with strong edge information $f_{ep} \in R^{C_{ep}\times h\times w}$ (See Fig. 4(c)). Unlike [8, 11] where parameters for edge detectors are fixed, all parameters in our edge-perceiving branch are learnable.

Considering that features from the semantic-perceiving branch f_{sp} and that from the edge-perceiving branch f_{ep} are responsible for perceiving different information and contributing unequally to parsing results, we follow [13] to fuse them a channel-wise attention method to exploit the relationship between the semantic information and the edge details. Lastly, the final predicted human parsing result is produced by a standard segmentation head.

4.4 Loss Functions

Human Parsing Loss. The typical cross-entropy loss is used for the final parsing results (L_{hp}) as well as the auxiliary supervision on the semantic-perceiving branch $\mathbf{S_{deep}}$ (L_{aux}).

Edge Perceiving Loss. Following [33], we adopt a weighted binary cross-entropy loss (L_{ep}) to supervise the training for edge-perceiving branch. The L_{ep} is defined as follows:

$$L_{ep} = -\sum_{i,c}\{1 : b_i > t\}\left(s_{i,c}\log\widehat{s_{i,c}}\right) \tag{1}$$

where t is the predefined threshold, b_i is the output of edge prediction head; $s_{i,c}$ and $\widehat{s_{i,c}}$ are the ground-truth and prediction result of the i-th pixel for class c, where $c \in (0,1)$. The ground-truth edge maps are generated following [27].

Fusion Loss. Features from the semantic-perceiving branch and the edge-perceiving branch will be fused and enhanced by the attention. Auxiliary supervisions are imposed on these enhanced features, deriving the fusion loss L_{fuse}:

$$L_{fuse} = L_{hp_{fuse}} + L_{ep_{fuse}} \tag{2}$$

Total Loss. The total loss L is a weighted combination of losses mentioned above:

$$L = \lambda_1 L_{hp} + \lambda_2 L_{ep} + \lambda_3 L_{fuse} \tag{3}$$

Note that all predicting heads but for the final segmentation head only work during the training phase.

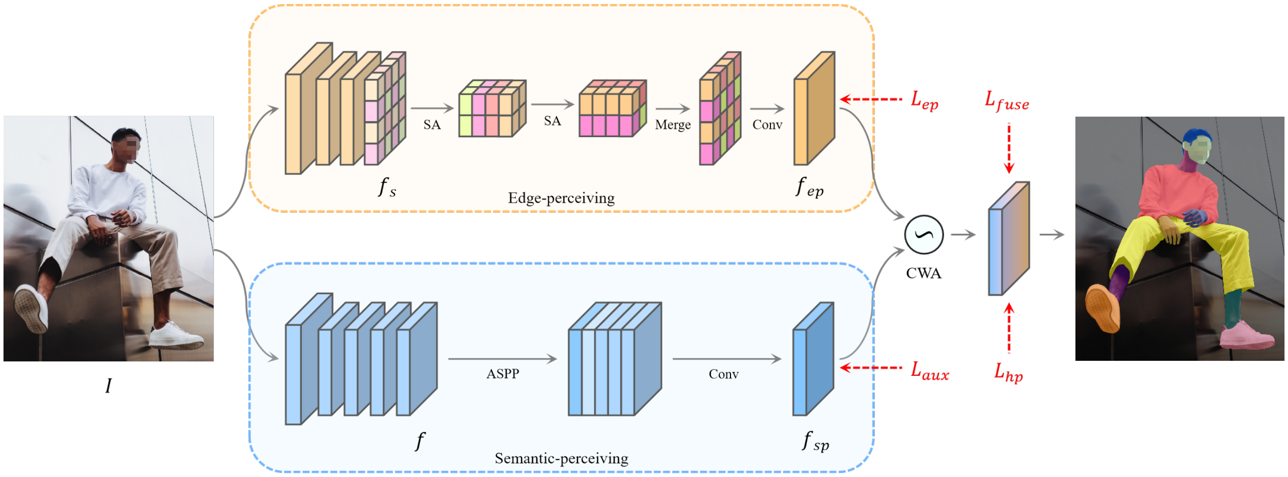
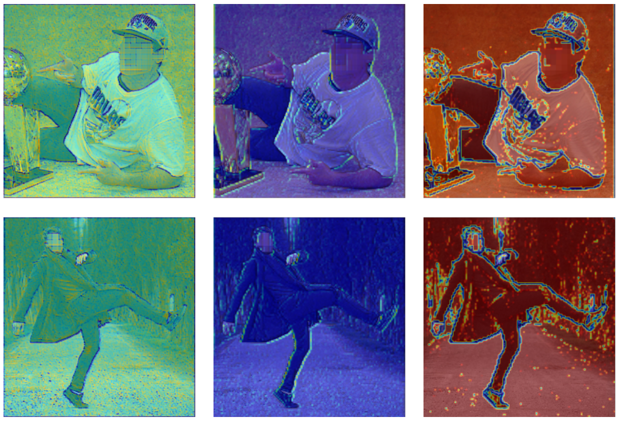


Figure 3: Pipeline of our proposed BiEPNet for high-resolution human parsing. The blue box is our semantic-perceiving branch to generate strong context information. The edge-perceiving branch shown in the yellow box provides edge details to refine details. Features are permuted before long-range and short-range self-attentions (**SA**) respectively. Features from two branches are fused with a channel-wise attention (**CWA**) module.



(a) DeepLabv3+ (b) BiSeNet (c) Ours

Figure 4: Comparison of the visualization of features in similar branches from DeepLabv3+, BiSeNet and our BiEPNet. The more red the parts are, the more contributions the parts make in the following steps. (a) is the visualized low-level features in DeepLabv3+ where details within closed regions rarely matter. (b) is the visualized features from the spatial path in BiSeNet where the edge information is weak. (c) is the visualized features from our edge-perceiving branch where pixels within a closed region are highly noticed, outlining more accurate edge information in turn.

5 Experiments

5.1 Implementation Details

Our method is composed of a semantic-perceiving branch and an edge-perceiving branch, where the semantic-perceiving is employed with DeepLabv3 [2] with ResNet18 [15] and the edge-perceiving branch consists of four convolution blocks by default. Parameters of ResNet18 are initialized byt the pretrained model on ImageNet [10]. Weights of all losses are set as 1.0 in Equation 3.

We apply the mmsegmentation [9] framework as our codebase. Data augmentations in human parsing are slightly different from generic semantic segmentation. Considering the memory usage during training and the continuity among the structures of the human body, we resize the longest edge of an input human image to 1024 while keeping its initial ratio. Next, we randomly scale the image with the scaling factor varied in (0.5,2). Furthermore, we crop a patch with 1024 × 1024 from the scaled image, balancing the continuity of human bodies and the abundance of training data. Noted that distinguishing the left/right pairs in human images is one of the special challenges, label pairs should be swapped if 'Flipping' occurs. Images will be padded into 1024 × 1024 during inference.

Parameters of our BiEPNet are optimized by SGD with momentum 0.9 and weight-decay $5e^{-4}$. The initial learning rate is set as 0.01 and we take a warmup strategy at the first 1K iterations with warmup factor 0.1. Then the learning rate is adjusted in a polynomial way with the decay parameter of 0.9 together with the maximum iterations set to 80K. Experiments are conducted with a batch size of 8 for training on 4 NVIDIA GeForce RTX 3080Ti GPUs. Memory usage is measured with batch size 1 on an RTX 3080Ti GPU.

5.2 Experiments on Human4K

We apply our framework to Human4K. Firstly, we compare our BiEP with several reproduced generic semantic segmentation methods on Human4K. Besides the standard evaluation metric mean intersection over union (mIoU), we also focus on the boundary performance, which is challenging in human parsing, averaging boundary IoU [7] for each class and acquiring **mean boundary IoU (mBIoU)**. For a fair comparison, all methods in Table 1 (except STDC [11]) take ResNet-18 [15] as the backbone and input sizes are 1024 × 1024.

In comparison to the approaches listed in Table 1, our method not only obtains the highest mIoU but also performs the best on boundary regions. Based on the quantitative result, there are two conclusions: 1) Benefiting from the design of the edge-perceiving branch, our BiEPNet outperforms on those hard categories, *e.g.* 'Socks', 'Left/Right pairs' and among which 'Scarf' is overwhelmingly better; 2) The semantic-perceiving branch yields a competitive understanding for those commonly seen categories, like 'Hair' and 'Pants'. We also list out the averaging memory consumption during inference for each method to demonstrate our efficiency. It is evident that CE2P is our main competitor in terms of performance, but it consumes 41.17% more memory than ours due to its redundant usage of low-level features from the backbone.

Fig. 5 gives the qualitative results and we mainly compare our methods with the baseline DeepLabv3+ and the state-of-the-art ISDNet. It can be found that DeepLabv3+ performs well in locating the subject in the image, but pixels within a semantic region could be wrongly classified (manifested as those spots in a region). Suffering

Table 1: Comparisons with some generic semantic segmentation models on Human4K validation set and IoU of some categories are given. The two best results are in red and blue. **mBIoU** is the averaged boundary IoU [7] for each class.

Methods	Hair	UpperClothes	Dress	Socks	Pants	Scarf	Left-shoe	Right-shoe	Torso-skin	mIoU	mBIoU	Mem(MB)
PSPNet [38]	79.71	62.32	53.87	**34.17**	77.3	**17.71**	46.6	47.2	72.45	54.86	34.67	5656
CE2P [27]	**81.89**	**64.74**	54.92	31.52	78.94	11.8	49.19	47.35	**74.74**	**56.45**	**36.62**	8048
PointRend [19]	78.9	60.51	47.01	32.37	75.58	13.96	48.48	50.11	72.92	54.23	35.28	7627
STDC [11]	80.52	64.07	53.69	31.06	**80.2**	8.96	50.2	**52.5**	74.25	56.37	36.56	4922
DeepLabv3+ [4]	81.07	**64.18**	**55.88**	33.82	**79.33**	11.53	**50.02**	49.87	**75.1**	56.42	36.23	8912
BiSeNetV1 [34]	78.65	62.51	54.73	29.63	77.13	11.54	50.03	51.9	71.81	55.31	34.89	5015
ISDNet [13]	77.38	61.76	54.77	23.54	75.36	9.42	48.8	49.22	67.78	53.78	33.37	4925
Ours	**81.32**	64.03	**57.4**	**35.36**	78.99	**39.85**	**50.46**	**52.37**	73.34	**57.94**	**36.76**	5701

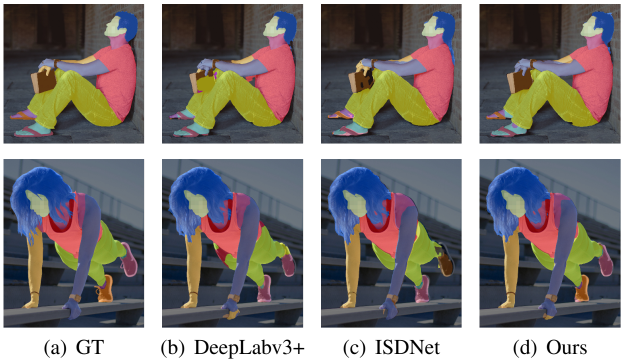

(a) GT (b) DeepLabv3+ (c) ISDNet (d) Ours

Figure 5: Comparison of qualitative results among our BiEP, baseline DeepLabv3+ and the state-of-the-art ISDNet. Our method presents stronger performance both on boundary parts and regions within a semantic region. 'L/R-Shoes' parts should be noticed.

Table 2: Comparison of results by different compositions. 'SP' represents the 'semantic-perceiving branch' and 'EP' represents the 'edge-perceiving branch'; w/o represents 'without' where self-attention is not adopted in the edge-perceiving branch and 'w' represents 'with'.

Composition	Acc	mACC	mIoU	mBIoU
SP	88.80	64.36	55.69	35.12
SP+EP(w/o SA)	89.68	67.99	57.20	36.57
SP+EP(w SA)	89.77	69.40	57.94	36.76

from the extreme downsampling in the deep branch, ISDNet presents a poor performance on details, leading to a blurred boundary recognition. Quantitative and qualitative experiments demonstrate that our BiEP successfully balances memory consumption and performance.

5.3 Ablation Studies

To illustrate the effectiveness of each module in our BiEPNet, we conduct a series of ablation studies, mainly focusing on our edge-perceiving branch.

Ablations of compositions. Our baseline method is the naive semantic-perceiving branch where DeepLabv3 with ResNet-18 is employed and the fusion method is the channel-wise attention by default. Table 2 shows that edge details provided by the edge-perceiving branch boost the overall performance by a large margin and the self-attention mechanism adopted on the top of the edge-perceiving branch further improves the performance.

Ablations of the number of convolution layers. Table 3 gives the results with different numbers of convolution layers in the edge-perceiving branch. When the number is smaller than three, the large feature maps will lead to running out of memory. The third line demonstrates that more convolution layers could ruin the overall performance, and thus we take four layers by default.

Table 3: Comparision of results with different numbers of convolution layers in the edge-perceiving branch.

# of Convs	Acc	mACC	mIoU	mBIoU
3	89.58	67.9	57.11	36.42
4	89.77	69.40	57.94	36.76
5	89.44	67.28	56.21	35.27

Ablations of fusion methods. Similar to [13], we compare different fusion methdos in Table 4. ADD is to directly add features that have the same size from the semantic-perceiving branch and the edge-perceiving respectively and pass a convolution block; while CAT is to concatenate the two features and pass a convolution block as well. Note that the latter method requires two times input channels as the former one. Comparison of the results demonstrates the effectiveness of the channel-wise attention fusion method.

Table 4: Comparision of results with different fusion methods.

Fusion Methods	Acc	mACC	mIoU	mBIoU
ADD	89.63	67.66	57.15	36.13
CAT	89.56	66.53	56.38	35.85
CWA	89.77	69.40	57.94	36.76

6 Conclusion and Limitations

This paper presents a high-resolution human parsing framework that consists of a semantic-perceiving branch and an edge-perceiving branch, endowing the network with superb abilities to process global and local information well at the same time. Two branches cooperate to contribute to the final parsing results through the channel-wise attention mechanism. Our memory-friendly method surpasses some state-of-the-art methods on our contributed high-resolution human parsing dataset Human4K.

However, due to the unbalance among categories, we notice that the training results could be varied with random seeds and this problem is not consistently resolved in this paper. A systematic exploration of more robust architectures and methods to preprocess human images is worth further research in the future.

ACKNOWLEDGMENTS

This work was supported in part by the National Key R&D Program of China (No.2021ZD0112100), National NSF of China (No.U1936212, No.62120106009).

REFERENCES

[1] J. Canny. A computational approach to edge detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, (6):679–698, 1986.

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

[3] L.-C. Chen, Y. Yang, J. Wang, W. Xu, and A. L. Yuille. Attention to scale: Scale-aware semantic image segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3640–3649, 2016.

[4] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *European Conference on Computer Vision*, pp. 801–818, 2018.

[5] W. Chen, Z. Jiang, Z. Wang, K. Cui, and X. Qian. Collaborative global-local networks for memory-efficient segmentation of ultra-high resolution images. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 8924–8933, 2019.

[6] B. Cheng, L.-C. Chen, Y. Wei, Y. Zhu, Z. Huang, J. Xiong, T. S. Huang, W.-M. Hwu, and H. Shi. Spgnet: Semantic prediction guidance for scene parsing. In *IEEE International Conference on Computer Vision*, pp. 5218–5228, 2019.

[7] B. Cheng, R. Girshick, P. Dollár, A. C. Berg, and A. Kirillov. Boundary iou: Improving object-centric image segmentation evaluation. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 15334–15342, 2021.

[8] H. K. Cheng, J. Chung, Y.-W. Tai, and C.-K. Tang. Cascadepsp: Toward class-agnostic and very high-resolution segmentation via global and local refinement. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 8890–8899, 2020.

[9] M. Contributors. MMSegmentation: Openmmlab semantic segmentation toolbox and benchmark. `https://github.com/open-mmlab/mmsegmentation`, 2020.

[10] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 248–255. Ieee, 2009.

[11] M. Fan, S. Lai, J. Huang, X. Wei, Z. Chai, J. Luo, and X. Wei. Rethinking bisenet for real-time semantic segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 9716–9725, 2021.

[12] K. Gong, X. Liang, D. Zhang, X. Shen, and L. Lin. Look into person: Self-supervised structure-sensitive learning and a new benchmark for human parsing. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 932–940, 2017.

[13] S. Guo, L. Liu, Z. Gan, Y. Wang, W. Zhang, C. Wang, G. Jiang, W. Zhang, R. Yi, and L. Ma. Isdnet: Integrating shallow and deep networks for efficient ultra-high resolution segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4361–4370, 2022.

[14] H. Harkat, J. Nascimento, and A. Bernardino. Fire segmentation using a deeplabv3+ architecture. In *Image and signal processing for remote sensing XXVI*, vol. 11533, pp. 134–145. SPIE, 2020.

[15] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *IEEE International Conference on Computer Vision*, pp. 770–778, 2016.

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

[17] C. Huynh, A. T. Tran, K. Luu, and M. Hoai. Progressive semantic segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 16755–16764, 2021.

[18] N. Kanopoulos, N. Vasanthavada, and R. L. Baker. Design of an image edge detection filter using the sobel operator. *IEEE Journal of solid-state circuits*, 23(2):358–367, 1988.

[19] A. Kirillov, Y. Wu, K. He, and R. Girshick. Pointrend: Image segmentation as rendering. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 9799–9808, 2020.

[20] P. Li, Y. Xu, Y. Wei, and Y. Yang. Self-correction for human parsing. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(6):3260–3271, 2020.

[21] X. Liang, S. Liu, X. Shen, J. Yang, L. Liu, J. Dong, L. Lin, and S. Yan. Deep human parsing with active template regression. *IEEE Transactions on Pattern Analysis Machine Intelligence*, 37(12):2402–2414, 2015.

[22] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft coco: Common objects in context. In *European Conference on Computer Vision*, pp. 740–755, 2014.

[23] K. Liu, O. Choi, J. Wang, and W. Hwang. Cdgnet: Class distribution guided network for human parsing. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4473–4482, 2022.

[24] J. Long, E. Shelhamer, and T. Darrell. Fully convolutional networks for semantic segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3431–3440, 2015.

[25] H. Polat. Multi-task semantic segmentation of ct images for covid-19 infections using deeplabv3+ based on dilated residual network. *Physical and Engineering Sciences in Medicine*, 45(2):443–455, 2022.

[26] 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.*, pp. 234–241. Springer, 2015.

[27] 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 Conference on Artificial Intelligence*, vol. 33, pp. 4814–4821, 2019.

[28] S. A. Taghanaki, A. Bentaieb, A. Sharma, S. K. Zhou, Y. Zheng, B. Georgescu, P. Sharma, Z. Xu, D. Comaniciu, and G. Hamarneh. Select, attend, and transfer: Light, learnable skip connections. In *Machine Learning in Medical Imaging*, pp. 417–425. Springer, 2019.

[29] C. Wang, P. Du, H. Wu, J. Li, C. Zhao, and H. Zhu. A cucumber leaf disease severity classification method based on the fusion of deeplabv3+ and u-net. *Computers and Electronics in Agriculture*, 189:106373, 2021.

[30] J. Wang, K. Sun, T. Cheng, B. Jiang, C. Deng, Y. Zhao, D. Liu, Y. Mu, M. Tan, and X. Wang. Deep high-resolution representation learning for visual recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(10):3349–3364, 2020.

[31] T. Wu, Z. Lei, B. Lin, C. Li, Y. Qu, and Y. Xie. Patch proposal network for fast semantic segmentation of high-resolution images. In *AAAI Conference on Artificial Intelligence*, pp. 12402–12409, 2020.

[32] C. Xie, C. Xia, M. Ma, Z. Zhao, X. Chen, and J. Li. Pyramid grafting network for one-stage high resolution saliency detection. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 11717–11726, 2022.

[33] J. Xu, Z. Xiong, and S. P. Bhattacharyya. Pidnet: A real-time semantic segmentation network inspired from pid controller, 2022.

[34] C. Yu, J. Wang, C. Peng, C. Gao, G. Yu, and N. Sang. Bisenet: Bilateral segmentation network for real-time semantic segmentation. In *European Conference on Computer Vision*, pp. 325–341, 2018.

[35] Y. Yuan, J. Xie, X. Chen, and J. Wang. Segfix: Model-agnostic boundary refinement for segmentation. In *European Conference on Computer Vision*, pp. 489–506, 2020.

[36] Z. Zhang, C. Su, L. Zheng, and X. Xie. Correlating edge, pose with parsing. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 8900–8909, 2020.

[37] Z. Zhang, C. Su, L. Zheng, X. Xie, and Y. Li. On the correlation among edge, pose and parsing. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(11):8492–8507, 2021.

[38] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia. Pyramid scene parsing network. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2881–2890, 2017.

[39] P. Zhou, B. Price, S. Cohen, G. Wilensky, and L. S. Davis. Deepstrip: High-resolution boundary refinement. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2020.