Image and Vision Computing 109 (2021) 104145

Contents lists available at ScienceDirect

Image and Vision Computing

journal homepage: www.elsevier.com/locate/imavis

FEANet: Foreground-edge-aware network with DenseASPOC for human parsing

Wing-Yin Yu [a,*], Lai-Man Po [a], Yuzhi Zhao [a], Yujia Zhang [a], Kin-Wai Lau [a,b]

[a] Department of Electrical Engineering, City University of Hong Kong, Hong Kong, China
[b] TCL Corporate Research Co., Ltd, Hong Kong, China

ARTICLE INFO

Article history:
Received 22 January 2021
Received in revised form 23 February 2021
Accepted 24 February 2021
Available online 01 March 2021

Keywords:
Human parsing
Semantic segmentation
Foreground-edge awareness
Non-local operation

ABSTRACT

Human parsing has drawn a lot of attention from the public due to its critical role in high-level computer vision applications. Recent works demonstrated the effectiveness of utilizing context module and additional information in improving the performance of human parsing. However, ambiguous objects, small scaling and occlusion problems are still the bottlenecks. In this paper, we propose a novel framework called - Foreground-Edge-Aware Network (FEANet) with DenseASPOC context module to further enhance the segmentation performance for human parsing. We claim that the fusion of foreground and edge information can effectively segment occluded regions by reducing the impact of pixels occupied by non-human object parts while persevering boundaries between each class. Moreover, we introduce the Dense Atrous Spatial Pyramid Object Context (DenseASPOC) module to address the problem of small and ambiguous objects by empowering feature extraction ability with solid spatial perception and semantic context information. We conducted comprehensive experiments on various human parsing benchmarks including both single-human and multi-human parsing. Both quantitative and qualitative results show that the proposed FEANet has superiority over the current methods. Moreover, detailed ablation studies report the effectiveness of the employment on each contribution.

1. Introduction

Human parsing deals with assigning a predefined label to each corresponding area of the human body such as hair, left-arm or dress, etc. Beneficial to this image-understanding-type task, many high-level computer vision applications can be realized such as clothing retrieval [1], clothing cosegmentation [2], object extraction for sports [3], fashion synthesis [4] and virtual try-on [5].

Semantic context information is an important component to boost segmentation performance by extracting semantic features for multi-scale objects. For example, DeepLab [6] proposed an atrous spatial pyramid pooling (ASPP) to utilize different rates of dilated convolution so that multi-scale features could be obtained by a large perceptive field of view. PSPNet [7] used the pyramid pooling module (PPM) to downscale and upscale different sub-region layers so that a comprehensive pooling representation can be concatenated. Although these context modules have achieved remarkable results in extracting semantic context information, their ability of extracting features on multi-scale objects (especially small-scale objects such as glove, socks, pants or scarf) is still weak. Examples can be viewed in Fig. 5(a)(b). To address these issues, we propose a new context module called Dense Atrous Spatial Pyramid Object Context (DenseASPOC) to fuse dense spatial information with object context knowledge. We claim that the dense connection among dilated convolutions accompanying with self-attention operation can effectively capture features for small-scale objects. It is because the dense dilated convolution is sensitive to geometric information. Self-attention operation can produce clues of object context. Such combination can consolidate the strength of feature extraction on multi-scale objects.

Apart from context module, using additional information as guidance is a common method to help network converge at an optimal point. For example, JPPNet [8] started cloth parsing work based on low-level image decomposition and pose landmark estimation. CE2P [9] established a new branch to predict the edge map for boundary enhancement. These guidance aims at remaining the characteristics of human body parts. However, the effectiveness is concerned when the non-human body parts overlap with the body. The occlusion effect hampers the correspondence among human body parts leading to unsatisfactory parsing result such as examples in Fig. 5(e)(f). Inspired by these works, we propose to filter out pixels occupied by non-human object parts by using the foreground information while preserving the boundaries by using the edge information. We believe that the fusion of foreground information and edge information can leverage a better

[*] Corresponding author.
E-mail addresses: wingyinyu8-c@my.cityu.edu.hk (W.-Y. Yu), eelmpo@cityu.edu.hk (L.-M. Po), yzzhao2-c@my.cityu.edu.hk (Y. Zhao), yzhang2383-c@my.cityu.edu.hk (Y. Zhang), kinwailau6-c@my.cityu.edu.hk (K.-W. Lau).

https://doi.org/10.1016/j.imavis.2021.104145
0262-8856/© 2021 Elsevier B.V. All rights reserved.

human parsing segmentation result by addressing the problem of occlusion and boundary separation.

In this paper, we propose an effective human parsing network called Foreground-Edge-Aware Network (FEANet) with a DenseASPOC context module to accomplish this challenging high-level computer vision task. The main idea of the proposed FEANet is to strengthen the ability to globally reduce the impact of pixels occupied by non-human object parts while locally preserving human part boundaries by using the foreground information and edge information. Specifically, we design a multi-purpose network that splits a common backbone into two branches including the foreground branch and edge branch. For the foreground branch, we take the advantages of spatial and semantic context module to capture the features of multi-scale objects, so that a fast-scanning purpose for the human part object can be achieved. For the edge branch, we focus on interpolating low-resolution feature maps accompanied with high-resolution maintenance. It aims to preserve the clear boundaries between each human part by embedding proportional semantic context information. Fig. 1 shows some examples of the additional guidance information used in our FEANet. For the DenseASPOC context module, it can consolidate the semantic meaning for certain objects by leveraging the relationship between geometric location and non-local information. It can enhance the ability of context feature extraction compared with existing context modules.

Our contributions can be summarized in the following three aspects:

- Fusion of foreground and edge information. We propose a novel network that uses foreground and edge information to enhance human parsing performance. This approach is to improve the global ability to reduce the impact of pixels occupied by non-human object parts while preserving the boundaries of local human object parts.
- DenseASPOC context module. We further study the method of extracting semantic features of multi-scale objects by combining geometric context with non-locally object-based context knowledge in a dense manner which can encapsulate deeper semantic context information.
- Experiments for single/multiple human parsing. We also conduct comprehensive experiments on several human parsing benchmarks demonstrating the generality of FEANet. The experimental results illustrate the effectiveness of FEANet for human parsing task. It achieves superior performance compared to the state-of-the-art approaches.

2. Related work

In this section, we give a brief survey of related works for semantic segmentation and human parsing domain.

2.1. Semantic segmentation

Fully Convolution Network (FCN) [10] was a prominent architecture for semantic segmentation by using deconvolution and fusion of pooling layers. Zhao et al. [7] developed a pyramid pooling module (PPM) in PSPNet to downscale and upscale different sub-region layers so that a comprehensive pooling representation could be concatenated. Similarly, Chen et al. [6] used different rates of dilated convolution to form an atrous spatial pyramid pooling (ASPP) module in DeepLab in order to enhance the context and details for feature perception. In addition to ASPP, Chen et al. [11] further combined the ASPP context module with encoder-decoder architecture in DeepLab to allow the network to capture features in both cross and intra layers. By extending ASPP [6], Yang et al. [12] proposed a DenseASPP module to assemble different dilated branches in a dense manner to solve the multi-scale objects problem by extensive global context information extraction. Although these

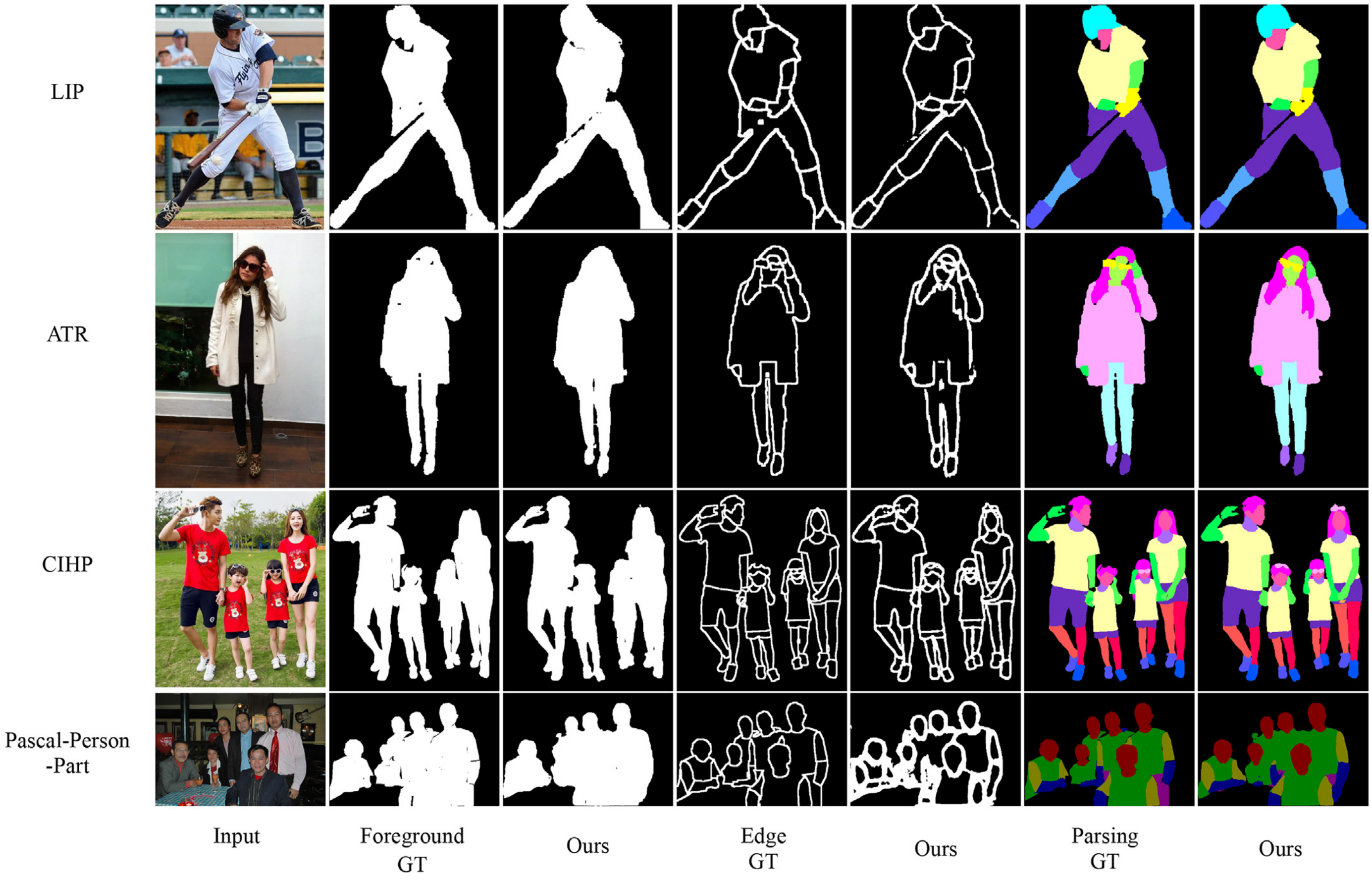


Fig. 1. It shows that not only does the FEANet correctly segment the final human parsing labels, but it also performs accurate foreground extraction and edge detection simultaneously.

state-of-the-art methods perform well in some semantic segmentation tasks such as scene parsing, they cannot be directly transferred to human parsing task. It is because the continuous downsampling operation by strided convolution and pooling method cause the problem of neglecting small objects and unclear boundaries between human parts which are two major challenges in human parsing.

2.2. Human parsing

Recently, some interventions utilized high-resolution representation. Liu et al. [13] proposed a braid module in BraidNet to exchange two-stream networks with low and high resolutions. It claimed that a model should learn high-level semantics from a deep yet narrow network while low-level spatial details from a shallow but wide network. With a similar hypothesis, Wang et al. [14] created a new backbone utilizing consistent high-resolution representation in the whole HRNet while interchanging low-resolution representation at the end of fusion sub-layers. HRNet [14] demonstrated that exchanging semantic information across different resolutions can preserve the detailed information for semantic segmentation. Last but not least, object-context-based methods had been applied to human parsing as well. For example, Yuan et al. [15] applied the object-context module in OCNet in order to exploit the object-context information to renovate object completeness via self-attention [16,17] techniques. Li et al. [18] demonstrated that applying non-local operation on the context module can effectively extract global information. Wang et al. [19] extended the object-context module in OCNet [15] to a comprehensive object context representation proposed in OCR by building a relationship between object representation and pixel representation via predicting soft-object regions in advance.

To put more emphasis on the boundary among objects, Ruan et al. [9] suggested splitting a network CE2P into two branches to jointly predict label map and edge map for the input image. It showed that edge information was an important factor to isolate human part objects with precise boundary enhancement. However, we believe that not only does the edge information help network to converge, the foreground information can be served as a guide to identify the non-human part regions. Therefore, we propose a method by fusing the edge and foreground information with the FEANet so that the occlusion issue can be solved. Moreover, the edge annotation used in CE2P [9] by computing the correlation of adjacent pixels cannot generate noise-free ground truths due to coarse label annotation quality. We also introduce a new method to generate a noise-free ground truth of edge map.

3. Proposed method

3.1. Problem definition

Human parsing, a sub-task of semantic segmentation particularly for performing analysis on human body parts, is defined to predict class labels for every pixel in a given image so that a class label map can be produced. Mathematically, the input image is an RGB image $\mathbf{I} \in \mathrm{R}^{W\times H\times \mathcal{C}}$ where it consists of dimension width W, height H and channel $\mathcal{C} = 3$. The output is a pixel-wise label map $\mathbf{M} \in \mathbb{R}^{W\times H}$ which has the same dimensions as the image $\mathbf{I}$. For each pixel value m_{xy} of label map $\mathbf{M}$, the corresponding value indicates the classified class $\mathbf{C} \in \{1,\ldots,c\}$ for the pixel i_{xy} of input image $\mathbf{I}$ where c is the total number of classes of interest.

3.2. Network architecture

The overview of the architecture for FEANet is shown in Fig. 2. The proposed FEANet is built based on the previous state-of-the-art method CE2P [9]. The network contains a ResNet-101 [20] as backbone which is widely adopted to extract sematic information in segmentation task. The final two feature maps namely Residual Block 4 and 5 of the backbone are passed to a context module composing a Densely connected Atrous Spatial Pyramid with an Object Context embedding operation. It is able to further exploit deep semantic low-level features in a fusion of geometric and object-based manner. The Foreground Aware Module (FAM) utilizes the feature from DenseASPOC and the shared feature Residual Block 2 from ResNet-101 [20] with skip connection to extract foreground features. Meanwhile, the Edge Aware Module (EAM) combines the multiple shared features including Residual Block 2, 3 and 4 from ResNet-101 [20] with a duplicated DenseASPOC module to perform edge detection. Finally, the Human Parsing Module (HPM) fuses all feature maps from FAM and EAM to refine the result of multi-class human parsing segmentation.

3.3. DenseASPOC context module

The objective of our DenseASPOC module is to strengthen the receptive field of the spatial context while retaining object context information. Inspired by [15], it followed similar intuition of self-attention approach and a densely connected atrous spatial pyramid pooling method. We advocate to enhance the ability of context extraction by densely combining geometric context with non-locally object-based context knowledge so that a correlation of geometric and semantic similarity information can be encapsulated. An illustration of DenseASPOC context module can be found in Fig. 3.

The purpose of object context pooling operation is to find out a pixel i belonging to one of the categories $\mathbf{C}$. We apply a non-local operation that computes the weighted mean of feature map to capture a long-range dependency from other references. Assuming the output feature layers from ResNet-101 [20] is $X \in \mathbb{R}^{W\times H\times F}$, we separately apply two convolution operations with an 1×1 kernel following a batch normalization [21] layer and a ReLU [22] activation layer to transform two distinct feature maps serving as the query and key representation. $\mathcal{Q} : \mathrm{R}^{W\times H\times F} \rightarrow \mathrm{R}^{W\times H\times Q}$ and $\mathcal{K} : \mathrm{R}^{W\times H\times F} \rightarrow \mathrm{R}^{W\times H\times K}$ are two feature

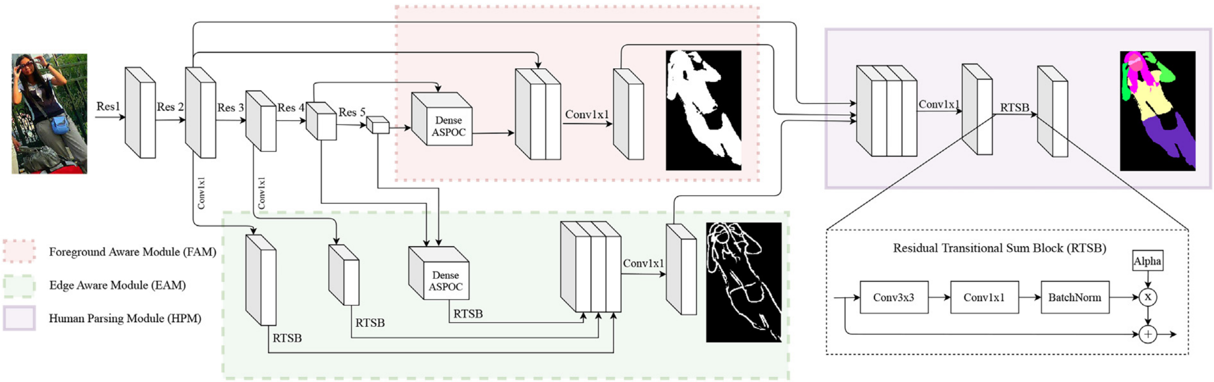
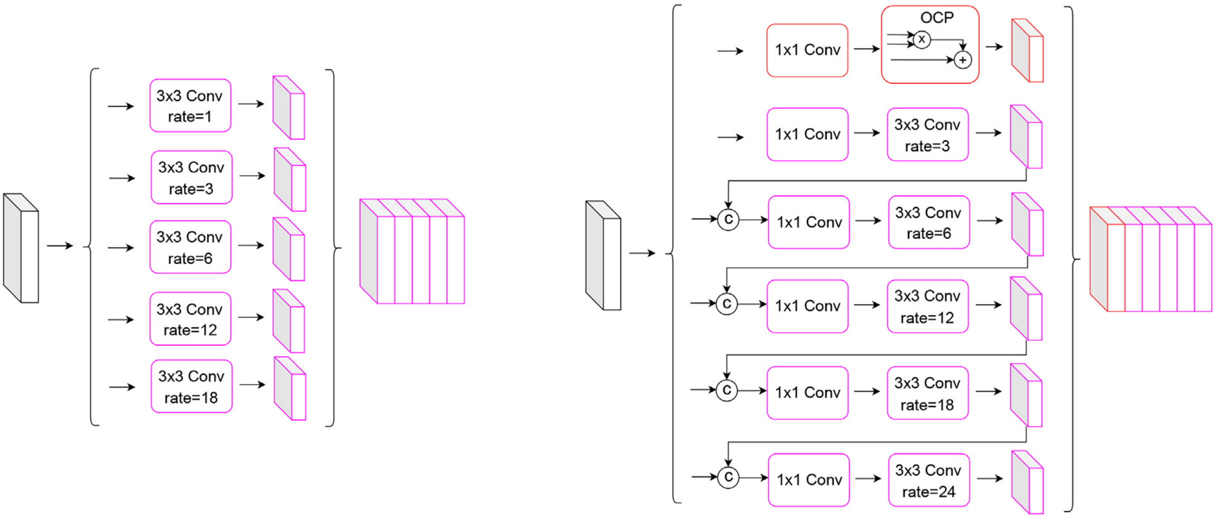


Fig. 2. Overview of the network architecture for the proposed FEANet.



(a) ASPP

(b) DenseASPOC (ours)

Fig. 3. Difference between various atrous spatial pyramid approaches. (a) Atrous Spatial Pyramid Pooling (ASPP) [6] module. (b) Dense Atrous Spatial Pyramid Object Context (DenseASPOC) module.

transformation functions where $Q = K$. Reshaping a three-dimensional matrix to two dimensional vectors for matrix multiplication, we can formulate the object context estimation representation $\mathscr{H} \in \mathbb{R}^{N \times N}$ as follows,

$$h_{ij} = \frac{exp\left(\mathcal{Q}_i \cdot \mathcal{K}_j^{\ T}\right)}{\sum_{n=1}^{N} exp\left(\mathcal{Q}_n \cdot \mathcal{K}_j^{\ T}\right)} \tag{1}$$

where $N = W \times H$ and normalizing with a soft-max activation function. Then, we aggregate context estimation representation with a value representation transformed from X. It can effectively establish the relationship between pixels with the object context. The value transformation function can be expressed as $\mathcal{V} : \mathrm{R}^{W \times H \times F} \rightarrow \mathrm{R}^{N \times V}$ combining with reshaping operation. The two-dimensional object context aggregation $\mathcal{O}\in\mathrm{R}^{N \times V}$ can be calculated by multiplying the object estimation representation with the value representation such as followings,

$$o_i = \sum_{j=1}^{N} h_{ij} \cdot \mathcal{V}_j \tag{2}$$

where $\mathcal{V}_j$ indicates the j^{th} row of categorical representation. After reversely reshaping $\mathcal{O}\in\mathrm{R}^{H \times W \times V}$, we can obtain the final object context representation.

Strengthening the ability feature extraction is another direction. We propose to make a dense connection among all the dilated convolutions within our context module. It creates a cascaded feature pyramid to strengthen the ability of feature extraction. Reviewing the traditional atrous convolution, it uses different dilated rates to expand the convolutional area while the dimension of feature map output is the same as input $\mathscr{H} \in \mathbb{R}^{W \times H \times G}$ which can be represented as followings,

$$\mathrm{H}_{K,d}[i] = \sum_{k=1}^{K} x[i + d \cdot k] \cdot w[k] \tag{3}$$

where K denotes the kernel size, d is the dilated rate and $w[k]$ is the k^{th} row of the kernel filter. Similar with ASPP [6] module, it concatenates all the features map from different dilated rates, it can be formulated as follows,

$$y =_{K,\ d \subset D} \mathrm{H}_{K,d}(X) \tag{4}$$

where D is a vector containing all dilated rates such as (3,6,12,18,24) and the ∥ operator indicates concatenation. Combining with the object context sub-module, it can be represented as below,

$$y = \mathcal{O}(X) \| \left(\mathop{\|}_{K,\ d \subset D} \mathrm{H}_{K,d}(X) \right) \tag{5}$$

where $\mathcal{O}$ is the object context pooling which has been elaborated above and the ∥ operator indicates concatenation. By making a dense connection, it stacks all the outputs of the dilated layers together to produce a multi-rate atrous convolution. The final structure of DenseASPOC is shown in Fig. 3(b). The formulation can be calculated as follows,

$$y = \mathcal{O}(X) \| \left(\mathop{\|}_{(D' \subset D)} \mathop{\|}_{(K, d \subset [D^1 \ldots D'])} \mathrm{H}_{K,d}(X) \right) \tag{6}$$

With such cascaded and parallel connection architecture, the DenseASPOC can further enlarge the receptive field of spatial context information preserving details of multi-scale shapes meanwhile it can yield a long-range dependency of object-based relationship through the self-attention operation.

3.4. Foreground/edge aware module

For the FAM, we take the advantages of spatial and semantic context module to capture the feature of multi-scale object so as to achieve the purpose of scanning human part objects. To maintain high-resolution characteristics, we concatenate the shared feature Residual Block 2 from the backbone with the output from the context module. It can perverse more spatial details during the encoding stage. By employing an 1 × 1 convolution, we can decode the feature layers to a two-channel foreground map.

For the EAM, we focus on interpolating low-resolution feature maps accompanied by the high-resolution maintenance while embedding a proportional semantic context information to preserve clear boundaries between each human part. We employ an 1 × 1 convolution with 256 filters and 3 × 3 based Residual Transitional Sum Block (RTSB) on the shared feature maps Residual Block 2, Residual Block 3 and the DenseASPOC context module. The RTSB is a simplified version of residual block in ResNet [23] and RefineNet [24]. In Fig. 2, it contains two consecutive convolutions with a 3 × 3 kernel while summarizing a weighted layer. It can help the network increase the generality and robustness which is certified in previous works [23,24]. The weight parameter alpha in RTSB is set to 0.1 referenced by the same setting as [24]. By fusing the three transitional blocks and employing an 1 × 1 convolution for decoding purpose, we can decode the feature layers to a two-channel edge map.

3.5. Human parsing module

With the aids of foreground and edge information, we can perform human parsing by fusing the predicted feature map from the FAM and the EAM. Unlike previous works [9], we make use of the last predicted feature map before softmax activation instead of three intermediate feature layers. For the final classification procedures, we apply a RTSB to ensure the robustness and a dropout layer with 0.1 dropout probability for regularization purpose. The total loss for FEANet can be formulated as follows,

$$\mathrm{L}_{total} = \ \lambda_1 \mathrm{L}_{fg} + \ \lambda_2 \mathrm{L}_{edge} + \ \lambda_3 \mathrm{L}_{parsing} + \ \lambda_4 \mathrm{L}_{IoU} \tag{7}$$

where $\mathscr{L}_{fg}$, $\mathscr{L}_{edge}$ and $\mathscr{L}_{parsing}$ represents the loss function implemented for the predicted foreground map, edge map prediction and human parsing prediction. $\mathscr{L}_{IoU}$ represents a loss dealing with indirect relationship with mean Intersection over Union (mIoU) metric. By making use of a tractable surrogate for the optimization of mIoU, we apply the Lov'asz-Softmax loss [25] to optimize the network with respect to mIoU metric. We adopt dual focal loss [26] which is an adaptive weighted loss for semantic segmentation to deal with the unbalanced-class problem. The objective functions for $\mathscr{L}_{fg}$, $\mathscr{L}_{edge}$ and $\mathscr{L}_{parsing}$ are as follows,

$$\mathrm{L}_{*} = \ -\sum_{i=1}^{N}\sum_{j=1}^{M} log\left(1-\|y_{i,j}-\widehat{y}_{i,j}\|_2^2\right) \tag{8}$$

where N and M represents the height and width of the predicted map, $\|\cdot\|_2^2$ represents L2 norm operator, $y_{i,\,j}$ are the predicted pixels along the channel vector on the feature map i^{th} row and j^{th} column, and $\widehat{y}_{i,j}$ are the pixels of the ground truth labels.

3.6. Ground truth enhancements

Previous edge map generation method, by computing the correlation of adjacent pixels [9], could effectively extract the edge information from the parsing label map. Due to coarse and noisy annotations from the datasets, such a generation approach produces discrete points on non-edge regions which were undesired for a ground truth label. Fig. 4 demonstrated the effect of discrete point produced by CE2P [9] and the enhancement of our method eliminating the noisy points.

To refine the edge map, we firstly apply morphology closing on the input label map $\mathbf{M}$ to fill up the disjoint regions. Secondly, we adopt the Canny edge detection method to extract the boundary of the refined label map $\mathbf{M}'$. Finally, using morphology dilation operation can increase the thickness of the edge pixels producing $\mathbf{y_e} \in \mathbb{R}^{W \times H}$. The operation procedures can be expressed by the following equations:

$$\mathbf{M}' = \mathbf{M} \cdot \mathbf{K_5} = (\mathbf{M} \oplus \mathbf{K_5}) \ominus \mathbf{K_5} \tag{9}$$

where $\mathbf{K_5}$ is a 5 × 5 kernel, ⊕ and ⊖ denotes dilation operation and erosion operation respectively.

$$\mathbf{y_e} = E(\mathbf{M}') = Canny(\mathbf{M}') \oplus \mathbf{K_3} \tag{10}$$

where $\mathbf{K_3}$ is a 3 × 3 kernel, $Canny(\cdot)$ represents the traditional Canny edge detection method.

For the ground truth of the foreground label, we simply binarize the refined label map $\mathbf{M}'$ to a two-channel semantic map $\mathbf{y_b} \in \mathbb{R}^{W \times H}$. The formulation can be calculated as follow:

$$\mathbf{y_f} = F(\mathbf{M}') = \begin{pmatrix} 1, & m'_{xy} \in \{\mathbf{C}\} \\ 0, & \text{otherwise} \end{pmatrix} \tag{11}$$

where $F(\cdot)$ represents the binarization operation, $m^{'}_{xy}$ indicates the pixels on the refined label map $\mathbf{M}^{'}$.

4. Experiments and results

In this section, we describe the implementation details for FEANet including datasets and pre-processing procedures. Moreover, we provide a comprehensive explanation of experimental results.

4.1. Datasets

LIP Dataset: The Look Into Person (LIP) [27] dataset is a large-scale benchmark widely used for single human parsing. The images are collected naturally from the wild where the people occur with different poses, viewpoints, lightings and occlusions in various real-world scenarios. It contains 50,462 images totally, including 30,462 for training, 10,000 for testing and 10,000 for validation. Every image comes with a fine-grained pixel-level annotation with 19 semantic human body part classes and one background class.

ATR Dataset: The Active Template Regression (ATR) [28] dataset is another commonly used dataset for single human parsing since 2015. The images are captured with limited conditions such as full-body especially for the commercial fashion design. There are 17,700 images totally, including 16,000 for training, 1000 for testing and 700 for validation. There are 18 human body part classes annotated with pixel-level labels.

PASCAL-Person-Part Dataset: It is another traditional challenging multi-human parsing dataset which is a sub-set of PASCAL-VOC 2010 [29] containing human-related samples only. There are 3533 images with fine-grained pixel-level annotations, including 1716 for training and 1817 for testing. By projecting class labels, there are 7 classes in total, for example, a background class and 6 human part labels.

CIHP Dataset: Comparing with LIP [27] dataset, the Crowd Instance-level Human Parsing (CIHP) [30] dataset is a large-scale benchmark for multi-human parsing task. There are 38,280 images totally, including 28,280 for training, 5000 for testing and 500 for validation. The annotation setting is the same as LIP [27] dataset which containing 20 categories in total including background.

4.2. Metrics

Following previous works [9,30–32], we report several evaluation metrics on different benchmarks which are defined as:

- mean Intersection over Union (mIoU): $({}^1\!/_C)\sum_i {}^{x_{ii}}\!/_{(p_i+\sum_j x_{ji}-x_{ii})}$,
- Pixel-wise Accuracy (Pix Acc.): $\sum_i {}^{x_{ii}}\!/_{\sum_i p_i}$,
- Mean Accuracy (Mean Acc.)/average precision (Prec.): $({}^1\!/_C)\sum_i {}^{x_{ii}}\!/_{p_i}$,
- Foreground Accuracy (F.G Acc.): $\sum_i {}^{x_{ii}}\!/_{\sum_i p_i}$, where $i \neq class_{background}$,
- Average Recall (Recall): $(1C)\sum_i {}^{x_{ii}}\!/_{\sum_j x_{ji}}$,
- Average F-1 (F-1): $({}^1\!/_C)\left({}^{MeanAcc. \times Recall}\!/_{(MeanAcc.+Recall)}\right)$,

where x_{ij} is the number of pixels of class i predicted to belong to class j, C is the total number of classes, and $p_i = \sum_j x_{ij}$ is the total number of pixels of class i.

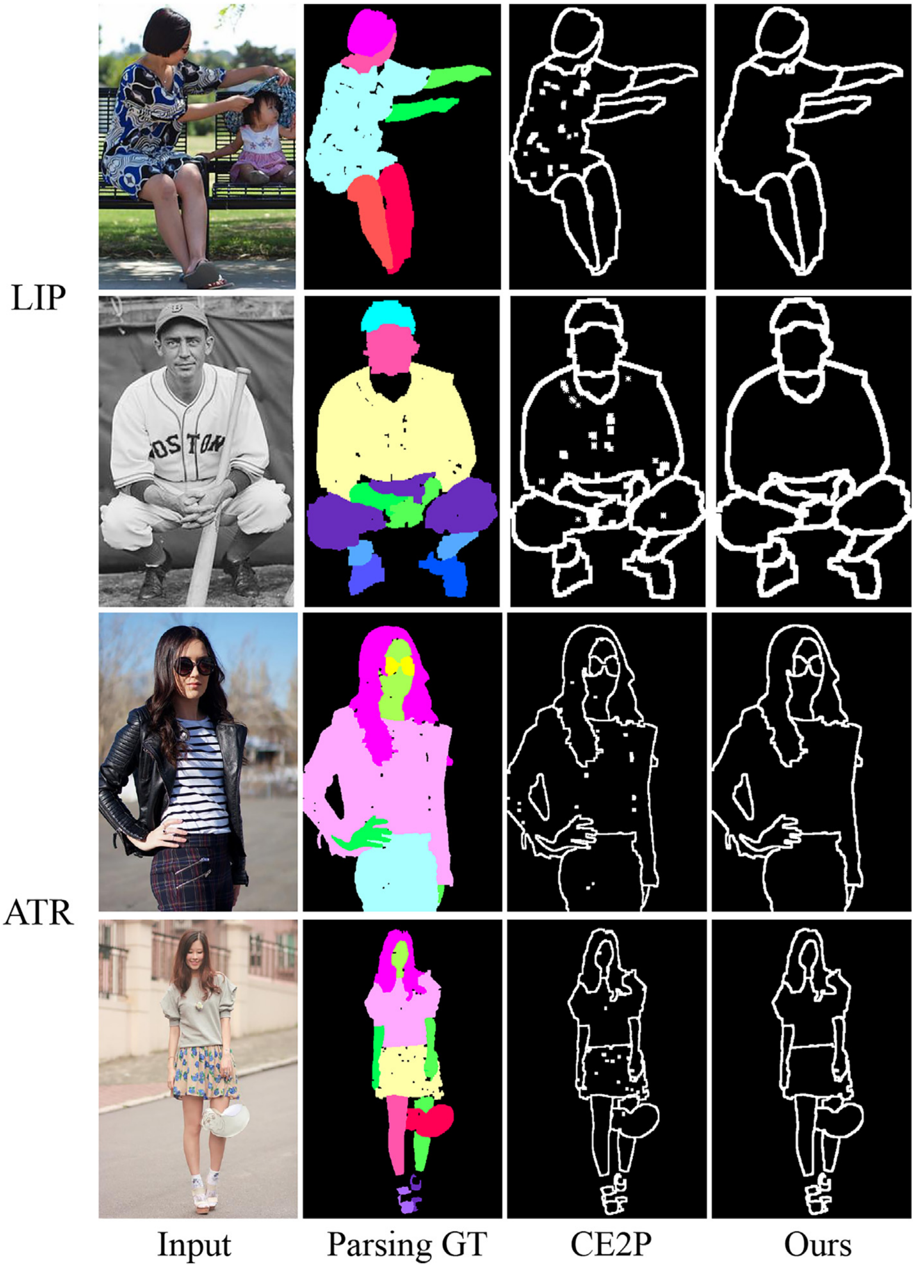


Fig. 4. Visualization of ground truth enhancement on edge map.

4.3. Pre-processing

To generalize the input size of image, we scale the size of input image to 473 × 473 for LIP [27] dataset, ATR [28] dataset and CIHP [30] dataset respectively, but 512 × 512 for PASCAL-Person-Part [29] during both training and testing. For augmentation part, we apply random scaling on the input image as well as parsing label ranging from 0.5 to 1.5. Accompanying with central cropping operation, we also flip left and right body parts (e.g. left/right hands or legs) to maximize the generality during inference, except PASCAL-Person-Part [29] dataset.

4.4. Experiment setting

We implement FEANet with the public framework PyTorch. Basically, we utilize ResNet-101 [23], pre-trained on ImageNet [33], as the backbone network. We apply a modified polynomial learning schedule policy $lr_{base}{}^{*}\left(1-\frac{iter}{epoch{*}len(dataloader)}\right)^{power}$ with the base learning rate at 0.01 and power at 0.9 for dual focal loss. The Stochastic Gradient Descent (SGD) optimizer is employed with mini-batch, momentum of 0.9 and weight decay of 0.0005 for training purpose. The weighting parameters λ_{*} are set to 1. All models are trained and tested on a GeForce RTX 2080 Ti GPU. The batch size is set to 5.

4.5. Experiment result on single-human parsing datasets

To illustrate the generality of our FEANet, we evaluate it on the single-human parsing datasets.

4.5.1. Performance on LIP dataset

As shown in Table. 1 and Table. 2, our FEANet outperforms the current state-of-the-art methods with promising improvement in

Table 1
Comparison of per-class IoU and mIoU with several state-of-the-art methods on validation set of LIP.

Method	hat	hair	glove	s. glass	u.clot	dress	coat	sock	pant	j.suit	scarf	skirt	face	l.arm	r.arm	l.leg	r.leg	l. shoe	r. shoe	bkg	mIoU
Attention [34]	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
DeepLab [6]	59.76	66.22	28.76	23.91	64.95	33.68	52.86	37.67	68.05	26.15	17.44	25.23	70.00	50.42	53.89	39.36	38.27	26.95	28.36	84.09	44.80
SSL [27]	58.21	67.17	31.20	23.65	63.66	28.31	52.35	39.58	69.40	28.61	13.70	22.52	74.84	52.83	55.67	48.22	47.49	31.80	29.97	84.64	46.19
MMAN [35]	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
JPPNet [8]	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 [9]	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
BraidNet [13]	66.80	72.00	42.50	32.10	69.80	33.70	57.40	49.00	74.90	32.40	19.30	27.00	74.90	65.50	67.90	60.20	59.60	47.40	47.90	88.00	54.40
FEANet(ours)	69.49	73.48	47.27	37.82	69.95	37.63	56.66	50.82	76.13	33.59	25.22	27.31	75.91	67.96	70.35	63.94	63.37	52.43	53.46	88.39	57.06

validation set of LIP [27] dataset. It gets the highest mIoU score of 57.06% and mean accuracy of 71.03%.

Comparing to the approaches based on DeepLabV3+ [6] including CE2P [9] (with PPM [7] module) and OCNet [15] (with ASPOC module), our FEANet gets a better mIoU score with 3.96% and 2.34% increment. Comparing to other methods such as BraidNet [13] and JPPNet [8], the proposed FEANet gets a significant improvement on every per-class IoU score. For those small-scale objects such as hat, hair, glove and socks, FEANet obtains improvement as well. Focusing on sunglass and scarf class, FEANets both yields over 5% IoU score enhancement. Moreover, some easily miss-classified and confusing class such left-shoe and right-shoe, FEANet also achieves over 5% gain which is a huge increment in human parsing task. It is beneficial to the role of object context and spatial information that can preserve such complex left–right semantic meaning. In general, our FEANet performs very well in LIP [27] dataset. It proves the effectiveness of human parsing ability when there is a complex scenario in the wild. The result can be interpreted as our DenseASPOC module can effectively perform semantic segmentation on human parsing by combining geometric context with non-locally object-based context knowledge to distinguish multi-scale objects.

4.5.2. Performance on ATR dataset

As shown in Table. 3, our FEANet significantly outperforms some state-of-the-art methods on ATR [28] dataset. Our method achieves the highest score in most of the evaluation metrics. More precisely, comparing to previous state-of-the-art approach TGPNet, FEANet gets over 1.8%, 4.7% and 3.3% improvement for average precision, average recall and average F-1 score. Moreover, our method surpasses ATR [28] method, which is the original maintainer of ATR [28] dataset, over 20% enhancement on average F-1 score. Furthermore, the FEANet performs better foreground segmentation than TGPNet. It is beneficial to the fusion of the foreground-aware module and edge-aware module of the proposed FEANet, which provides assistance for the separation of multi-class object parts. It illustrates that the foreground information can provide positive effect for human parsing task. The evaluation result supports that our FEANet is effective on full human body parsing because full body setting is the main characteristic of ATR [28] dataset.

Table 2
Comparison of human parsing performance with several state-of-the-art methods on validation set of LIP.

Method	Pix Acc.	Mean Acc.	mIoU
Attention [34]	83.43	54.39	42.92
Attention+SSL [27]	84.36	54.94	44.73
MMAN [35]	85.24	57.60	46.93
SS-NAN [36]	87.59	56.03	47.92
HSP-PRI [37]	85.07	60.54	48.16
MuLA [38]	88.50	60.50	49.30
PSPNet [7]	86.23	61.33	50.56
CE2P [9]	87.37	63.20	53.10
BraidNet [13]	87.60	66.09	54.42
FEANet (ours)	87.95	71.03	57.06

Table 3
Comparison of human parsing performance with several state-of-the-art methods on test set of ATR.

Method	Pix Acc.	F.G Acc.	Prec.	Recall	F-1
Yamaguchi [39]	84.38	55.59	37.54	51.05	41.80
Paperdoll [40]	88.96	62.18	52.75	49.43	44.76
M-CNN [41]	89.57	73.98	64.56	65.17	62.81
ATR [28]	91.11	71.04	71.69	60.25	64.38
PSPNet [7]	95.20	80.23	79.66	73.79	75.84
Attention [34]	95.41	85.71	81.30	73.55	77.23
DeepLabV3+ [6]	95.96	83.04	80.41	78.79	79.49
Co-CNN [28]	96.02	83.57	84.95	77.66	80.14
LS-LSTM [42]	96.18	84.79	84.64	79.43	80.97
TGPNet [31]	96.45	87.91	83.36	80.22	81.76
FEANet (ours)	96.63	87.98	85.24	84.93	85.08

4.6. Experiment result on multi-human parsing datasets

Similar with single-human parsing, we compare our work with several state-of-the-art approaches on some multi-human parsing benchmarks.

4.6.1. Performance on PASCAL-person-part dataset

In Table. 4, our FEANet gets competitive performance over the current state-of-the-arts. Comparing with WSHP [49], our FEANet gets a little improvement on mIoU score but superior enhancement (3.83%) on both upper-arm and low-arm classes. However, there are rooms for improvement in our FEANet to achieve more gains in mIoU score comparing with previous state-of-the-art PGN [30] which using extra instance

Table 4
Comparison of human parsing performance with several state-of-the-art methods on test set of Pascal-Person-Part.

Method	head	torso	u.arm	l.arm	u.leg	l.leg	bkg	mIoU
HAZN [43]	80.79	59.11	43.05	42.76	38.99	34.46	93.59	56.11
Attention [34]	81.47	59.06	44.15	42.50	38.28	35.62	93.65	56.39
LG-LSTM [42]	82.72	60.99	45.40	47.76	42.33	37.96	88.63	57.97
Att + MMAN [35]	82.58	62.83	48.49	47.37	42.80	40.40	94.82	59.91
GraphLSTM [44]	82.69	62.68	46.88	47.71	45.66	40.93	94.59	60.16
Struc.LSTM [45]	82.89	67.15	51.42	48.72	51.72	45.91	97.18	63.57
MuLA [38]	84.60	68.30	57.50	54.10	49.60	46.40	95.60	65.10
PCNet [46]	86.81	69.06	55.35	55.27	50.21	48.54	96.07	65.90
Holistic [47]	86.00	69.85	56.63	55.92	51.46	48.82	95.73	66.30
SAN [48]	86.12	73.49	59.20	56.20	51.39	49.58	96.01	67.42
WSHP [49]	87.15	72.28	57.07	56.21	52.43	50.36	97.72	67.60
PGN [30]	90.89	75.12	55.83	64.61	55.42	51.70	95.33	68.40
FEANet (ours)	86.22	69.54	60.90	60.04	51.57	49.57	95.53	67.62

information. We believe that the instance information such as bounding box is effective in multi-human parsing so the PGN [30] yields a better performance. On the other hand, the FEANet may suffer from insufficient generality of small-scale dataset that containing around 1700 training samples and 7 classes only. It limits the ability of the dense perception from DenseASPOC. It is implied that our FEANet relies on comprehensive varieties of training samples in order to maximize the effectiveness of the foreground and edge modules.

4.6.2. Performance on CIHP dataset

As shown in Table. 5, our FEANet yields superior performance comparing with other state-of-the-art solutions in terms of mean accuracy and mIoU score. In particular, FEANet gets around 1.3% and 1.8% improvement on mIoU score over Parsing R-CNN [32] and BraidNet [13]. Noted that both of these two methods utilize extra information from object detection task while our FEANet uses pixel-level semantic labels only. Moreover, comparing with PGN [30], our method obtains 6.6% enhancement on mIoU score and 10.3% improvement on mean accuracy. This significant enhancement can be beneficial to the large amount of training data since CIHP [30] includes much more data samples (28,280 for training) and more prediction classes (20 classes). Although such comprehensive generality of dataset increases difficulty of human parsing task, it provides more perception clues for our DenseASPOC module to distinguish the semantic meaning during inferencing. In addition, the foreground and edge modules can effectively alleviate the impact of pixels occupied by non-human object parts while preserving local human part boundaries. It can also interpret that our FEANet can outperform some methods with object detection assistance when it satisfies the condition that there are sufficient training samples. The result and analysis are also matched along with PASCAL-Person-Part [29] dataset.

4.7. Ablation study

To prove the effectiveness and robustness of our FEANet, we conduct a series of experiments on most of our contributions. We evaluate all the settings on the LIP [27] dataset with mIoU score in Table. 6. Generally, the full model of FEANet gains the highest mIoU score compared to other variants of FEANet. It can be interpreted that all the contributions of FEANet get positive effect towards an accurate human parsing segmentation performance. The details of each setting are described as following.

4.7.1. CE2P

Our FEANet is built upon the previous state-of-the-art method CE2P [9] which demonstrated remarkable result on human parsing. Comparing to this approach, our FEANet outperforms 3.96% mIoU score which is a huge improvement. With the same PPM [7] context module, the FEANet still surpasses 2.95% mIoU score. By applying the proposed ground truth enhancement method, an improvement of more than 0.5% can be obtained. In order to show our contribution fairly, we evaluated each proposed component on this baseline, including DenseASPOC, foreground module, the enhanced edge module, and RTSB. The results in Table. 6 show that most of the proposed methods can independently provide positive gain to the baseline, and the enhanced edge module can achieve better performance of 1.08%. Although the RTSB module cannot improve the baseline performance, when it is applied to the proposed FEANet, the effect is considerable (+ 0.67%). We believe that the weighting parameter alpha used in RTSB is a variant of the entire network architecture.

Table 5
Comparison of human parsing performance with several state-of-the-art methods on validation set of CIHP.

Method	Mean Acc.	mIoU
PGN [30]	64.22	55.80
DeepLabV3 + [11]	65.06	57.13
BraidNet + MaskRCNN [13]	–	60.62
Parsing R-CNN [32]	–	61.10
FEANet (ours)	74.55	62.48

Table 6
Ablation Study – Comparison of human parsing performance with different components.

Method	mIoU
Analysis on the baseline	
CE2P [9] (baseline)	53.10
CE2P [9] with gt enhancement	53.64
CE2P [9] with DenseASPOC	53.16
CE2P [9] with fg	53.12
CE2P [9] with enhanced edge	54.18
CE2P [9] with RTSB	51.11
Analysis on additional information	
FEANet w/o fg	56.89
FEANet w/o edge	56.65
FEANet w/o fg and edge	56.80
Analysis on context modules	
FEANet with PPM [7]	56.05
FEANet with ASPP [6]	56.07
FEANet with ASPOC [15]	56.47
FEANet with DenseASPP [12]	56.74
DenseASPOC w/o dense connection	56.04
DenseASPOC with dilated rates = {12,24,36}	56.05
DenseASPOC w/o prior convolution	56.74
FEANet w/o gt enhancement	56.26
FEANet w/o RTSB	56.39
FEANet full model	57.06

4.7.2. Analysis on additional information

To evaluate the effectiveness of extra-information assistance we compare three settings including FEANet without foreground, without edge and without both of them. It means that there is no gradient backward from $\mathscr{L}_{fg}$ and $\mathscr{L}_{edge}$ respectively in Eq. (7) during training. As shown in Table. 6, the mIoU scores of the three settings drop from 0.17% to 0.41%. More significantly, there are a lot of decrements for the FEANet without the assistance of edge map since the edge map can effectively help declare the boundaries among each human part object. For foreground information, FEANet hardly helps to further improve the overall performance of human parsing, because it can provide some clues to reduce the impact of pixels occupied by non-human object parts.

4.7.3. Analysis on context modules

Apart from additional information, we also provide quantitative results on FEANet with different context modules including PPM [7], ASPP [6], ASPOC [15] and DenseASPP [12]. In Table. 6, it is clear that our FEANet with DenseASPOC module (i.e. full model) obtains the highest mIoU score compared to other context modules. Specifically, our context module surpasses around 1% mIoU score when comparing with PPM [7] and ASPP [6] module. There are also 0.59% and 0.32% score enhancement on ASPOC [15] and DenseASPP [12] module. It is realized that the non-local operation can help network to perceive object-based information from a long range of relationship. In human parsing, it provides clues from other human part objects rather than just surrounding pixels. In order to compare the performance of DenseASPOC and ASPOC in [15], we provide the evaluation results of three settings, such as no dense connections, larger dilated rates and no prior convolution. As shown in Table. 6, the full model can increase by at least 0.3% compared to its variants. We observe that, compared with ASPOC [15], dense connections within DenseASPOC can enhance the network by more than 1%. This shows that our DenseASPOC module can perform more accurate and robust semantic segmentation on human parsing tasks compared with existing context modules.

4.7.4. FEANet without ground truth enhancement

We replace the ground truth of edge map to the one generated from CE2P [9] method which computing the correlation of adjacent pixels. As shown in Table. 6, our algorithm on edge map generation can improve 0.8% compared to the CE2P [9] method. It is beneficial from noise removal on the original label map so that a clean boundary extraction can be executed. The increment on mIoU score can be certified the positive effect on our ground truth enhancement algorithm on edge map.

4.7.5. FEANet without RTSB

In term of network architecture, we further make enhancement on the conventional convolution with the 3 × 3 kernel. Regarding to the sub-block – RTSB, the evaluation results show that the FEANet enquired with RTSB can get positive gains (0.67%) within the network compared to original design. It is realized that it can maintain details from previous layers while extracting deeper features for referencing.

4.8. Qualitative analysis

We also show some examples of edge map and foreground map predicted from FEANet while proving their effectiveness. Moreover, experiments on label comparison and heat map are qualitatively conducted in this section as well.

4.8.1. Comparison with the state-of-the-arts

Fig. 5 demonstrates the visualizations on different state-of-the-art approaches such as MMAN [35], JPPNet [8] and CE2P [9] on LIP [27] dataset. The descriptions of each method can be referred to Section 2 Related Work. Overall, our FEANet achieves the best parsing visualization performance in the following three circumstances.

Firstly, the FEANet can accurately capture small objects of human part which are difficult to segment. For instance, some small objects including socks in Fig. 5(a), glove in Fig. 5(a)(c) and sunglass in Fig. 5(b)(c)(d)(f) are correctly segmented in our FEANet module, where the parsing ability on small objects is aligned with the quantitative experiments as well. It means that our FEANet can obtain good segmentation performance on human parts with small scale. It is beneficial from the role of DenseASPOC module contributing hints from long range dependency and comprehensive perception of the surroundings.

Secondly, the FEANet is effective on some easily confusing classes such as scarf. In Fig. 5(c), we can observe that scarf object is usually misclassified as upper-cloth or coat since the texture and position are similar. It is clear that our FEANet can correctly segment the scarf part on Fig. 5(c) while it achieves the highest 25.22% mIoU score compared to other methods. It is the consequence of object context relationship from our DenseASPOC module proving non-local dependency with other objects.

Last but not least, our FEANet performs well in the situation of intensive occlusions and boundaries. In Fig. 5(e)(f), the tennis racket and the advertisement board overlap some parts of human body which increasing the difficulty of human parsing task. It requires a clear cut from foreground (human part objects) with background (non-human part objects). The result shows that the FEANet can handle occlusion and boundary problems very well, where the tennis racket and the board are perfectly segmented as background class while maintaining an accurate segmentation on upper-clothes, pants and dress, for example.

4.8.2. Visualization on additional information

We provide a qualitative visualization on various dataset regarding edge map, foreground map and label parsing. In Fig. 1, we demonstrate

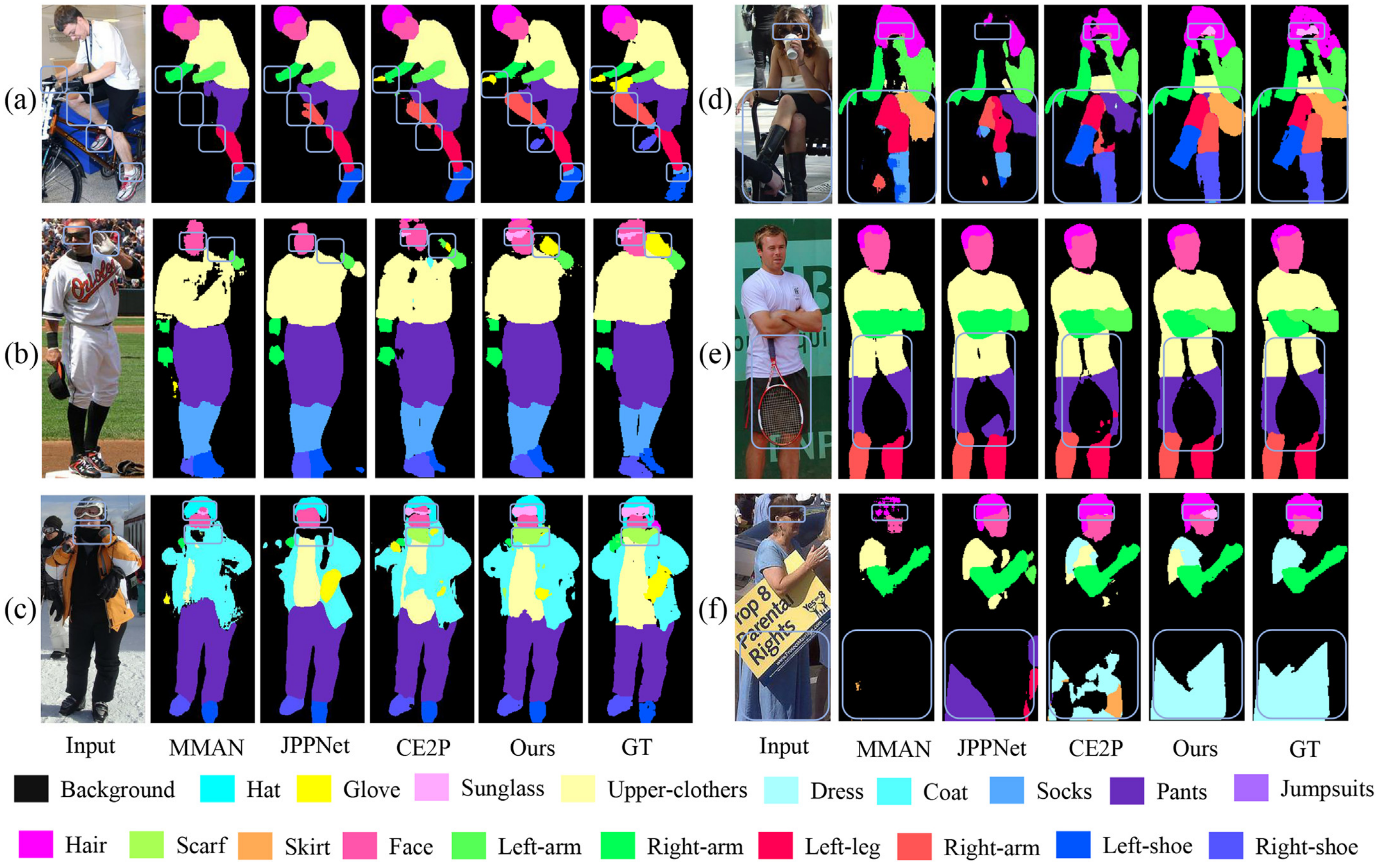


Fig. 5. Visual comparison of parsing performance on different state-of-the-art methods.

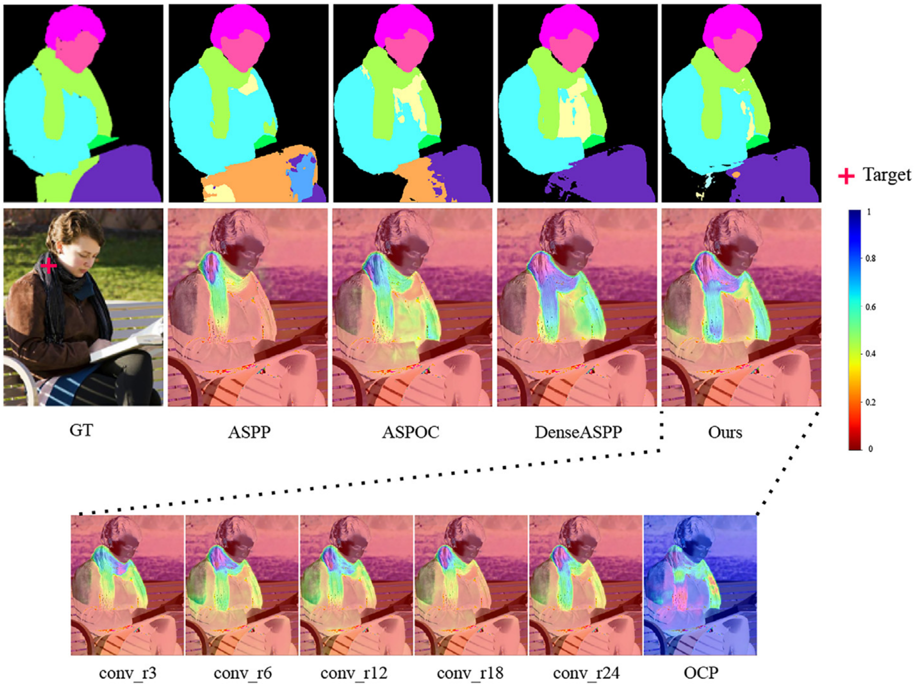


Fig. 6. Visual comparison of receptive fields of different well-known context modules. Red marking indicates the reference pixel. The spectrum starts from red to purple representing the least confidence to the most confidence. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

some examples on LIP [27], ATR [28], PASCAL-Person-Part [29] and CIHP [30] dataset. Compared to the ground truths of foreground map and edge map, our FEANet can perform well in foreground and edge detection in both single and multiple human parsing. It can separate the human part objects from the wild scene. With embedding boundary preserving ability, the FEANet can clearly partition the legs object into three parts including pants, socks and shoes in LIP [27] dataset; coat and jumpsuits in ATR [28] dataset; sunglass and hair in CIHP [30] dataset and upper-arms and lower-arms in PASCAL-Person-Part [29] dataset.

4.8.3. Visualization on different context modules

To illustrate the effectiveness of the proposed DenseASPOC module, we provide visualizations on intermediate results of different context modules since they are all related to atrous convolution. For a fair comparison, we simply replace the context module units from DenseASPOC to others while keeping consistent setting of FEANet during the whole experiment. Specifically, we collect the final feature block from context module which containing 512 feature maps in total. To obtain the target score on a particular class (e.g. scarf in Fig. 6), we forward an input image multiplied with every normalized feature map. It can produce a weighted heat map showing confidence scores to a reference pixel. Moreover, we visualize every feature map computed by atrous convolution with different dilated rates and the object context pooling module. The spectrum starts from red to purple representing the least confidence to the most confidence.

In Fig. 6, we show the heat maps produced from each context module on the targeted scarf class (indicated by a red marking on the input image). In consequence, our DenseASPOC module can produce the most accurate distribution on the pixels possibly belonging to the target class. It leads to an accurate segmentation result. From the result of atrous convolution with different dilated rates, we can observe that the densely connected atrous convolution helps locally capture features from a large region of surrounding pixels to highlight the possible pixel regions. Meanwhile, from the result of object context pooling sub-module, the module tries to obtain semantic meanings from long range dependency by non-local operation. The perceptive field can be expanded to the whole image so as to build connections with other objects. The result demonstrates that it can successfully filter out target-class objects such as scarf in this example. Therefore, it can prove the ability of feature perception and global dependency of DenseASPOC.

5. Conclusion

This paper proposes FEANet to solve the problem of ambiguous small objects, and it performs well under occlusion conditions. We propose using foreground and edge information to enhance human parsing performance by improving the ability to globally reduce the impact of pixels occupied by non-human object parts while preserving the boundaries of human parts locally. We further studied the method of extracting semantic features of multi-scale objects by combining geometric context with non-locally object-based context knowledge in a dense manner which can encapsulate deeper semantic context information. Both quantitative and qualitative experimental results have proved the effectiveness of our foreground-aware module, edge-aware module and the most important DenseASPOC context module. It can also achieve the state-of-the-art performance in both single/multi-human parsing benchmarks which demonstrating the promising robustness and generality in human parsing task.

CRediT authorship contribution statement

Wing-Yin Yu: Conceptualization, Software, Writing - original draft. **Lai-Man Po**: Supervision, Writing - review & editing. **Yuzhi Zhao**: Visualization, Writing - review & editing. **Yujia Zhang**: Investigation, Writing - review & editing. **Kin-Wai Lau**: Validation, Writing - review & editing.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] X. Liang, L. Lin, W. Yang, P. Luo, J. Huang, S. Yan, Clothes co-parsing via joint image segmentation and labeling with application to clothing retrieval, IEEE Trans. Multimedia 18 (6) (2016) 1175–1186.

[2] B. Zhao, X. Wu, Q. Peng, S. Yan, Clothing cosegmentation for shopping images with cluttered background, IEEE Trans. Multimedia 18 (6) (2016) 1111–1123.

[3] Z. Lou, T. Gevers, Extracting primary objects by video co-segmentation, IEEE Trans. Multimedia 16 (8) (2014) 2110–2117.

[4] S. Zhu, R. Urtasun, S. Fidler, D. Lin, C. Change Loy, Be your own prada: Fashion synthesis with structural coherence, Proceedings of the IEEE International Conference on Computer Vision 2017, pp. 1680–1688.

[5] M. Leo, P. Carcagnì, P.L. Mazzeo, P. Spagnolo, D. Cazzato, C. Distante, Analysis of facial information for healthcare applications: a survey on computer vision-based approaches, Information 11 (3) (2020) 128.

[6] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, A.L. Yuille, Deeplab: semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs, IEEE Trans. Pattern Anal. Mach. Intell. 40 (4) (2017) 834–848.

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

[8] X. Liang, K. Gong, X. Shen, L. Lin, Look into person: joint body parsing & pose estimation network and a new benchmark, IEEE Trans. Pattern Anal. Mach. Intell. 41 (4) (2018) 871–885.

[9] T. Ruan, T. Liu, Z. Huang, Y. Wei, S. Wei, Y. Zhao, Devil in the details: Towards accurate single and multiple human parsing, Proceedings of the AAAI Conference on Artificial Intelligence, vol. 33, 2019, pp. 4814–4821.

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

[11] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, H. Adam, Encoder-decoder with atrous separable convolution for semantic image segmentation, Proceedings of the European Conference on Computer Vision (ECCV) 2018, pp. 801–818.

[12] M. Yang, K. Yu, C. Zhang, Z. Li, K. Yang, Denseaspp for semantic segmentation in street scenes, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition 2018, pp. 3684–3692.

[13] X. Liu, M. Zhang, W. Liu, J. Song, T. Mei, Braidnet: Braiding semantics and details for accurate human parsing, Proceedings of the 27th ACM International Conference on Multimedia 2019, pp. 338–346.

[14] J. Wang, K. Sun, T. Cheng, B. Jiang, C. Deng, Y. Zhao, D. Liu, Y. Mu, M. Tan, X. Wang, et al., Deep high-resolution representation learning for visual recognition, IEEE Trans. Pattern Anal. Mach. Intell. (2020)https://doi.org/10.1109/TPAMI.2020.2983686.

[15] Y. Yuan, J. Wang, Ocnet: Object Context Network for Scene Parsing, arXiv preprint arXiv:1809.00916 2018.

[16] Z. Lin, M. Feng, C.N.D. Santos, M. Yu, B. Xiang, B. Zhou, Y. Bengio, A Structured Self-Attentive Sentence Embedding, arXiv preprint arXiv:1703.03130 2017.

[17] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, L. Kaiser, I. Polosukhin, Attention is all you need, Advances in Neural Information Processing Systems 2017, pp. 5998–6008.

[18] T. Li, W. Wan, Y. Huang, J. Chen, C. Hu, Y. Ma, Improving human parsing by extracting global information using the non-local operation, 2019 IEEE International Conference on Image Processing (ICIP), IEEE 2019, pp. 2961–2965.

[19] Y. Yuan, X. Chen, J. Wang, Object-Contextual Representations for Semantic Segmentation, arXiv preprint arXiv:1909.11065 2019.

[20] A. Krizhevsky, I. Sutskever, G.E. Hinton, Imagenet classification with deep convolutional neural networks, Advances in Neural Information Processing Systems 2012, pp. 1097–1105.

[21] S. Ioffe, C. Szegedy, Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift, arXiv preprint arXiv:1502.03167 2015.

[22] V. Nair, G.E. Hinton, Rectified linear units improve restricted boltzmann machines, ICML, 2010.

[23] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition 2016, pp. 770–778.

[24] G. Lin, A. Milan, C. Shen, I. Reid, Refinenet: Multi-path refinement networks for high-resolution semantic segmentation, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition 2017, pp. 1925–1934.

[25] M. Berman, A. Rannen Triki, M.B. Blaschko, The lovász-softmax loss: A tractable surrogate for the optimization of the intersection-over-union measure in neural networks, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition 2018, pp. 4413–4421.

[26] M.S. Hossain, A.P. Paplinski, J.M. Betts, Adaptive Class Weight Based Dual Focal Loss for Improved Semantic Segmentation, arXiv preprint arXiv:1909.11932 2019.

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

[28] X. Liang, S. Liu, X. Shen, J. Yang, L. Liu, J. Dong, L. Lin, S. Yan, Deep human parsing with active template regression, IEEE Trans. Pattern Anal. Mach. Intell. 37 (12) (2015) 2402–2414.

[29] X. Chen, R. Mottaghi, X. Liu, S. Fidler, R. Urtasun, A. Yuille, Detect what you can: Detecting and representing objects using holistic models and body parts, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition 2014, pp. 1971–1978.

[30] K. Gong, X. Liang, Y. Li, Y. Chen, M. Yang, L. Lin, Instance-level human parsing via part grouping network, Proceedings of the European Conference on Computer Vision (ECCV) 2018, pp. 770–785.

[31] X. Luo, Z. Su, J. Guo, G. Zhang, X. He, Trusted guidance pyramid network for human parsing, Proceedings of the 26th ACM International Conference on Multimedia 2018, pp. 654–662.

[32] L. Yang, Q. Song, Z. Wang, M. Jiang, Parsing r-cnn for instance-level human analysis, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition 2019, pp. 364–373.

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

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

[35] Y. Luo, Z. Zheng, L. Zheng, T. Guan, J. Yu, Y. Yang, Macro-micro adversarial network for human parsing, Proceedings of the European Conference on Computer Vision (ECCV) 2018, pp. 418–434.

[36] J. Zhao, J. Li, X. Nie, F. Zhao, Y. Chen, Z. Wang, J. Feng, S. Yan, Self-supervised neural aggregation networks for human parsing, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops 2017, pp. 7–15.

[37] M.M. Kalayeh, E. Basaran, M. Gökmen, M.E. Kamasak, M. Shah, Human semantic parsing for person re-identification, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition 2018, pp. 1062–1071.

[38] X. Nie, J. Feng, S. Yan, Mutual learning to adapt for joint human parsing and pose estimation, Proceedings of the European Conference on Computer Vision (ECCV) 2018, pp. 502–517.

[39] K. Yamaguchi, M.H. Kiapour, L.E. Ortiz, T.L. Berg, Parsing clothing in fashion photographs, 2012 IEEE Conference on Computer Vision and Pattern Recognition, IEEE 2012, pp. 3570–3577.

[40] K. Yamaguchi, M. Hadi Kiapour, T.L. Berg, Paper doll parsing: Retrieving similar styles to parse clothing items, Proceedings of the IEEE International Conference on Computer Vision 2013, pp. 3519–3526.

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

[42] X. Liang, X. Shen, D. Xiang, J. Feng, L. Lin, S. Yan, Semantic object parsing with local-global long short-term memory, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition 2016, pp. 3185–3193.

[43] F. Xia, P. Wang, L.-C. Chen, A.L. Yuille, Zoom better to see clearer: Human and object parsing with hierarchical auto-zoom net, European Conference on Computer Vision, Springer 2016, pp. 648–663.

[44] X. Liang, X. Shen, J. Feng, L. Lin, S. Yan, Semantic object parsing with graph lstm, European Conference on Computer Vision, Springer 2016, pp. 125–143.

[45] X. Liang, L. Lin, X. Shen, J. Feng, S. Yan, E.P. Xing, Interpretable structure-evolving lstm, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition 2017, pp. 1010–1019.

[46] B. Zhu, Y. Chen, M. Tang, J. Wang, Progressive cognitive human parsing, Thirty-Second AAAI Conference on Artificial Intelligence, 2018.

[47] Q. Li, A. Arnab, P.H. Torr, Holistic, Instance-Level Human Parsing, arXiv preprint arXiv:1709.03612 2017.

[48] X. Huang, K. Wu, G. Hu, J. Shao, Multi-class human body parsing with edge-enhancement network, International Conference on Neural Information Processing, Springer 2019, pp. 466–477.

[49] H.-S. Fang, G. Lu, X. Fang, J. Xie, Y.-W. Tai, C. Lu, Weakly and Semi Supervised Human Body Part Parsing Via Pose-Guided Knowledge Transfer, arXiv preprint arXiv:1805.04310 2018.