BraidNet: Braiding Semantics and Details for Accurate Human Parsing

Xinchen Liu[1], Meng Zhang[1], Wu Liu[1*], Jingkuan Song[2], Tao Mei[1]
[1]AI Research of JD.com, [2]University of Electronic Science and Technology of China
liuxinchen1@jd.com,zhangmeng10@bupt.edu.cn,liuwu@live.cn,jingkuan.song@gmail.com,tmei@live.com

ABSTRACT

This paper focuses on fine-grained human parsing in images. This is a very challenging task due to the diverse person appearance, semantic ambiguity of different body parts and clothing, and extremely small parsing targets. Although existing approaches can achieve significant improvement by pyramid feature learning, multi-level supervision, and joint learning with pose estimation, human parsing is still far from being solved. Different from existing approaches, we propose a Braiding Network, named as BraidNet, to learn complementary semantics and details for fine-grained human parsing. The BraidNet contains a two-stream braid-like architecture. The first stream is a semantic abstracting net with a deep yet narrow structure which can learn semantic knowledge by a hierarchy of fully convolution layers to overcome the challenges of diverse person appearance. To capture low-level details of small targets, the detail-preserving net is designed to exploit a shallow yet wide network without down-sampling, which can retain sufficient local structures for small objects. Moreover, we design a group of braiding modules across the two sub-nets, by which complementary information can be exchanged during end-to-end training. Besides, in the end of BraidNet, a Pairwise Hard Region Embedding strategy is propose to eliminate the semantic ambiguity of different body parts and clothing. Extensive experiments show that the proposed BraidNet achieves better performance than the state-of-the-art methods for fine-grained human parsing.

CCS CONCEPTS

• **Computing methodologies** → **Computer vision**; **Image segmentation**; **Neural networks**.

KEYWORDS

Fine-grained Human Parsing, Semantic Segmentation, Braiding Network, Pairwise Hard Region Embedding

ACM Reference Format:
Xinchen Liu, Meng Zhang, Wu Liu, Jingkuan Song, and Tao Mei. 2019. BraidNet: Braiding Semantics and Details for Accurate Human Parsing. In *Proceedings of the 27th ACM International Conference on Multimedia (MM'19), October 21–25, 2019, Nice, France.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3343031.3350857

*Wu Liu is the corresponding author.

MM '19, October 21–25, 2019, Nice, France

ACM ISBN 978-1-4503-6889-6/19/10…$15.00
https://doi.org/10.1145/3343031.3350857

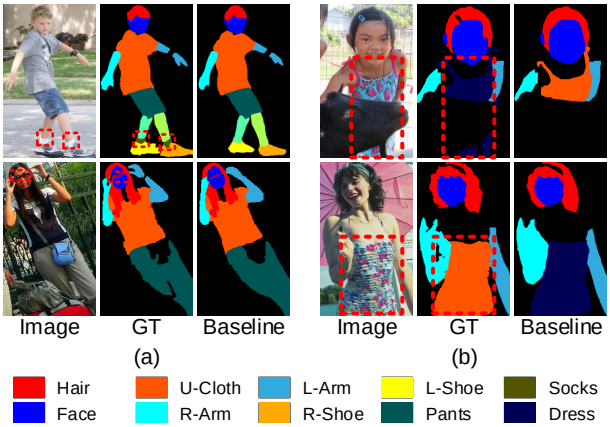

Figure 1: The challenges of human parsing. (a) The extremely small targets. (b) The ambiguous clothes that have similar visual appearance. (Best viewed in color.)

1 INTRODUCTION

Human parsing, also known as fine-grained human part segmentation, is a fundamental task in multimedia and computer vision. Human parsing is the task of pixel-level classification on human parts or clothes, e.g., hair, face, dress, etc. It has attracted tremendous attention from the community due to its potential applications such as fashion analysis [21], person re-identification [14], event detection [8], and image translation [6, 30]. However, the variety of person appearance and similarity among parsing targets make human parsing a very challenging task which needs further exploration.

As a specific image segmentation task, existing human parsing methods usually adopt the fully convolutional networks (FCNs) [27] or encoder-decoder networks [1] for semantic segmentation as the basic models [10, 18]. For semantic segmentation, researchers exploit atrous convolution to obtain multi-scale field-of-views on images [38] and perform pyramid feature fusion to capture multi-level information in FCNs [40], which significantly improve the performance for semantic segmentation as well as human parsing. Moreover, some researchers consider human parsing as a mask generation problem and adopt generative adversarial networks (GANs) to obtain accurate parsing results [29, 41]. In addition, recent studies adopt multi-task learning to exploit edges [9, 22] or key points [17] to model the structures of body parts and clothes.

However, existing methods for human parsing usually neglect two challenges. The first is that the small targets may be missed, such as the sunglasses shown in Figure 1(a). The reason is that, as the feature maps are down-sampled through the FCNs, small targets and details are overwhelmed by surrounding large targets or the background. Besides, the extremely unbalanced distribution of targets also makes the learning-based methods, i.e., FCNs, prefer large objects. The second is the ambiguity among different body parts and clothes due to the visual similarity of appearance, variations of poses or viewpoints, and occlusions. For example, as shown in Figure 1(b), the upper-clothes and dress have similar texture and shape that even humans can hardly distinguish them.

To overcome the above challenges, we propose a Braiding Network, dubbed as BraidNet, for fine-grained human parsing. The BraidNet contains a two-stream braid-like fully convolutional architecture, which can learn both high-level semantics and detailed structures from images. In particular, it has a semantic abstracting net with a deep yet narrow network which exploits a bottom-up hierarchy to abstract semantic knowledge.[1] To capture low-level details lost in the bottom-up path, we design a detail-preserving net which has a shallow and wide structure without down-sampling to retain sufficient local structures for small targets. Moreover, we design a group of braiding modules to connect the layers in the two sub-nets. The braiding modules work as a pipeline by which the two networks can exchange complementary information during end-to-end training.

Furthermore, we propose a Pairwise Hard Region Embedding (PHRE) strategy for training the BraidNet to distinguish similar targets. In other areas of multimedia, e.g., fine-grained image recognition [39], pairwise metric learning is adopted to learn a latent space to discriminate similar objects. However, the image-level metric learning cannot be directly adopted into human parsing which outputs the pixel-level dense prediction, since it is inefficient and difficult to sample sufficient pixel pairs to learn the discriminative metric space. Moreover, the pixel-level features can hardly provide high-level semantic knowledge to represent the parsing targets. Therefore, in PHRE, we first sample ambiguous region pairs based on a graph which models the ambiguity relationships among different targets. For a pair of images with ambiguous regions, we propose a hard-aware regional representation to measure their distance in the metric space. During training, the PHRE can pull samples of the same class closer while making samples of different classes scattered, especially for the ambiguous classes. Finally, the BraidNet is optimized with both pixel-level supervision and regional metric learning for human parsing.

In summary, the contributions of this paper include:

- We propose a Braiding Network with two parallel sub-nets for human parsing. One sub-net is a deep and narrow-down network to learn semantic knowledge. The other is a shallow but wide network to capture local structures from images.
- To effectively explore semantics and local structures, we design the braiding module to exchange information between the two sub-nets, which makes the BraidNet learn robust and discriminative features, especially for small targets.
- We design a PHRE strategy which can make the BraidNet learn to differentiate ambiguous parsing targets through a hard-aware regional metric learning scheme.

Through extensive experiments on two public benchmarks, the proposed BraidNet outperforms the state-of-the-art methods.

[1] In this paper, the "narrow" and "wide" indicate the size of feature maps from convolution or pooling layers, instead of the convolution kernel numbers as other papers.

2 RELATED WORK

Semantic Segmentation. Semantic segmentation has achieved significant improvement since convolutional neural networks (CNNs), especially FCNs, are explored to learn a pixel-to-pixel mapping from large-scale data [27, 31]. Recent FCN-based methods achieve excellent results for multi-scale objects, especially small targets, by well-designed convolution kernels [4, 38] and/or integrating multi-scale features from one layer or cross multiple layers in FCNs [2, 3, 40]. For example, Yu and Koltun proposed the atrous convolution to learn robust features for varied scales [38]. Dai *et al.* proposed a deformable convolutional network with deformable kernels to learn effective features for objects with varied shapes and scales [4]. Zhao *et al.* proposed a pyramid pooling module to aggregate contextual information with different scales and achieved excellent results on scene parsing [40]. Chen *et al.* proposed an atrous spatial pyramid pooling (ASPP) module to apply multi-scale atrous convolution on one feature map, which could capture both details and context in one layer [2]. Chen *et al.* integrated the encoder-decoder network with the ASPP module to combine multi-level features in both intra-layer and cross-layer manners and obtained the state-of-the-art performance on semantic segmentation [3]. Although above methods provide principles of designing networks for human parsing, these methods only perform multi-scale feature learning in one single FCN. The detailed structures and small targets could be still overwhelmed by hierarchical down-sampling due to the convolutions with stride larger than one and the pooling operations. Therefore, we propose a two-stream framework which has one FCN for semantic context and the other FCN for local details and small targets. By a series of braiding modules, the two networks can exchange complementary information and learn a discriminative representation for human parsing.

Human Parsing. Although human parsing is a specific task of semantic segmentation, researchers have delved into the unique characteristics of human body and clothes, and proposed elaborated models for both single human parsing [5, 17, 18, 28, 29, 36, 37] and instance-level multiple human parsing [9, 22, 41, 42]. For example, in the early years, researchers proposed to parse clothes by template matching and retrieval based approaches [36, 37]. The key points of human body were explored to provide structural information for human parsing [17, 37]. Recently, Luo *et al.* adopted a generative adversarial network (GAN) with a macro discriminator and a micro discriminator to make the generator output robust parsing results [29]. Zhao *et al.* proposed a hierarchy of three GANs to generate foreground, instance, and body parts for instance-level human parsing.[41]. Gong *et al.* proposed a part grouping network to simultaneously generate the parsing masks and edges of persons for instance-level human parsing[28]. Luo *et al.* proposed a trusted guidance pyramid network supervised by multi-level supervision during training [29]. Liu *et al.* proposed to jointly generate the masks and edges of parsing targets by multi-task learning, which

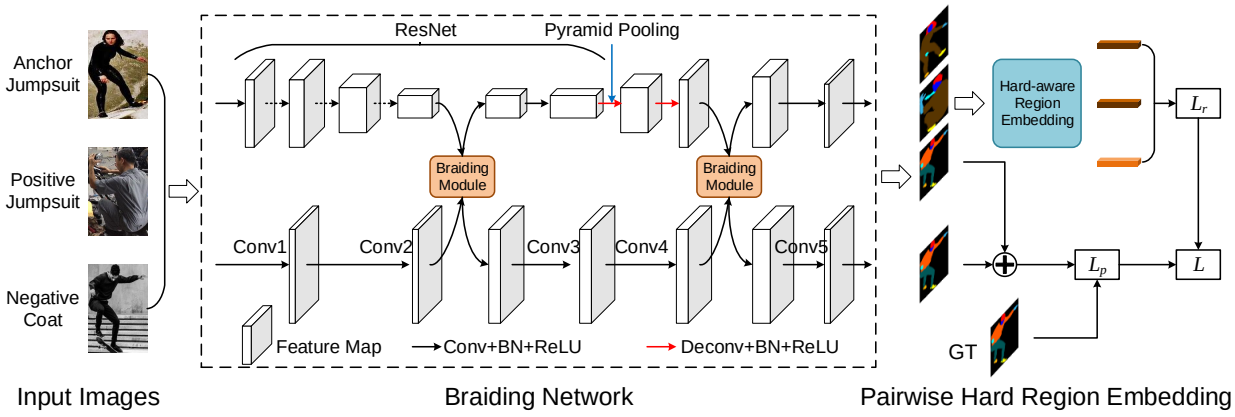

Figure 2: The overall architecture of the Braiding Network with Pairwise Hard Region Embedding. The upper network is the semantic abstracting net. The lower one is the detail-preserving net. (Best viewed in color.)

achieved the state-of-the-art results for human parsing[22]. Nevertheless, these methods neglect the ambiguity among different body parts and clothes due to their appearance similarity, varied posed, and occlusion. Therefore, we propose the Pairwise Hard Region Embedding for BraidNet to learn discriminative representations and eliminate the ambiguity in fine-grained human parsing.

3 APPROACH

3.1 Overview

The overall architecture of the Braiding Network with Pairwise Hard Region Embedding is shown in Figure 2. The BraidNet has two FCN-based networks: the semantics abstracting net and the detail-preserving net. The former has a narrow-down architecture as regular FCNs to learn high-level semantics, i.e, the classes of objects, from raw pixels. While the latter is a wide FCN without down-sampling to preserve local structures. To effectively explore the semantics and details, we design the Braiding Modules between the two networks, which can make them exchange information during training. Moreover, the BraidNet is optimized by a novel Pairwise Hard Region Embedding strategy which makes the network learn to differentiate ambiguous targets. Next, we present the detailed structures of the BraidNet and the Braiding Module, then introduce the PHRE strategy.

3.2 Braiding Network For Human Parsing

Semantic Abstracting Net. The global semantics of targets is important to guarantee complete and continuous parsing results. For example, we can easily recognize a complete T-shirt rather than only a small path of cloth. To accurately segment a fine-grained part, the model should know the shape, position, and surrounding from the global view. As shown in Figure 2, we adopt the PSPNet [40] as the semantic abstracting net in our BraidNet, since the pyramid pooling module (PPM) of PSPNet can effectively aggregate multi-scale semantic context. Moreover, the PSPNet has a concise structure without other branches, which makes the framework easy for optimization. As in [40], we adopt the ResNet-101 [12] pretrained on ImageNet dataset [34] as the backbone. The last convolution stage of the backbone is connected with a four-level PPM which has pooling kernels of 1×1, 2×2, 3×3, and 6×6, respectively. At last, one deconvolution layer and one convolution layer are adopted to decode the semantics and output the dense prediction. Through the hierarchical narrow-down structure of ResNet and pyramid pooling on the high-level feature maps, the semantic information is abstracted from images for human parsing.

Table 1: The architecture for the detail-preserving net.

Layer Name	Output Size	Kernel Size & Number
Conv1	192×192	7×7, 64
Conv2	192×192	3×3, 128
Conv3	192×192	3×3, 128
Conv4	192×192	3×3, 256
Conv5	192×192	3×3, 20

Detail-Preserving Net. In conventional CNNs or FCNs, the size of input images and feature maps shrink during forward propagation due to the convolution with stride larger than one and the pooling operations. This may make the detailed structures, such as the edge, blurred and the small targets like shoes and sunglasses overwhelmed. To overcome this problem, existing methods usually adopt encoder-decoder architecture which directly concatenates the low-level feature maps to the high-level ones [22, 38]. However, the multi-level features are still learned by the same network and should be further explored. Therefore, we design a detail-preserving net without down-sampling to guarantee high-resolution of feature maps. The detail-preserving net contains five convolution layers as shown in Figure 2. At the tail of the network, the softmax operation is adopted to obtain the dense prediction. The parameters of each layer are listed in Table 1. The strides of all convolution layers are set to 1 to keep the size of feature maps fixed.

Braiding Module. As mentioned above, the semantic abstracting net and the detail-preserving net are adopted to learn semantic context and local structures, respectively. How to effectively exploit these knowledge is still a problem. One straight-forward way is to

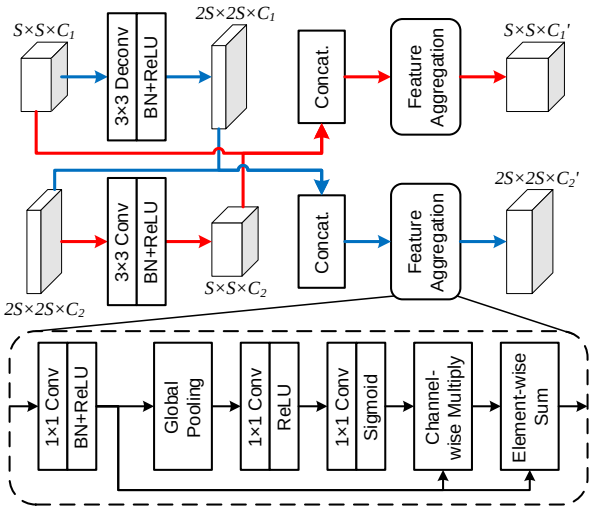

Figure 3: The structure of braiding module.

combine the results of the two networks by late fusion. Another scheme is by mid fusion which integrates the last feature maps of the two networks and output one parsing result from the fused feature. However these two strategies cannot effectively discovery the complementary information. Therefore, we design the braiding module between the two networks, as shown in Figure 2.

The structure of the braiding module is shown in Figure 3. There are two data streams in the braiding module. In the first stream (the blue line in Figure 3), the $S \times S \times C_1$ feature map from the semantic abstracting net is up-sampled to $2S \times 2S \times C_1$ by a deconvolution layer with stride=2. Then the up-sampled feature map is concatenated with the $2S \times 2S \times C_2$ feature map from the detail-preserving net to obtain a $2S \times 2S \times (C_1 + C_2)$ one. At last, through a three-layer channel attention operation, we obtain a $2S \times 2S \times C_2'$ feature map as the input of the next convolution layer in the detail-preserving net. The second stream (the red line in Figure 3) has similar operation except that the feature map from the detail-preserving net is first down-sampled by a convolution layer with stride=2, and its output is a $S \times S \times C_1'$ feature map for the next layer of the semantic abstracting net. Theoretically, the braiding module can be inserted between any pair of convolution layers in the two networks. Because we aim to exchange high-level semantics and local details between the two networks. In our implementation, we insert one braiding module between the last two convolution layers of the two networks. The other braiding module is embedded between the Conv3 and Conv4 in the semantic abstracting net, and between Conv2 and Conv4 in the detail-preserving net, as shown in Figure 2.

In summary, our Braiding Network has a semantic abstracting net to learn high-level knowledge about the parsing targets and a detail-preserving net that is focused on local structures and small objects. With the braiding module, the complementary information is exchanged between the two networks. Therefore, the multi-level features can be effectively aggregated for accurate human parsing.

3.3 Pairwise Hard Region Embedding

As discussed in Section 1, one of the main challenges of fine-grained human parsing is the similarity among parsing targets due to the ambiguity of objects under different viewpoints, poses, and occlusions. Typical examples include left and right arms/legs/shoes, coat and upper-clothes, skirt and pants, etc. In other areas of multimedia and computer vision, such as image retrieval [25, 39], face recognition [23, 35], zero-shot learning [7], and object re-identification [13, 24], feature embedding by metric learning has been widely adopted to solve the problems of the intra-class difference and the inter-class similarity. For semantic segmentation, Kohl *et al.* proposed a probabilistic U-Net to segment lung abnormality images [15]. Rui *et al.* proposed a point-based distance metric learning method to segment images with only a few point annotations [32]. Existing metric learning methods usually apply constraints on the distance between features of image pairs to learn an embedding space in which the inter-class distance is much larger than the intra-class distance. However, this scheme cannot be directly adopted to human parsing, since human parsing has pixel-level output and label. It is difficult and inefficient to sample sufficient pixel pairs during training. Moreover, pixel-level features contain little semantic knowledge to differentiate similar targets. Therefore, we propose a Pairwise Hard Region Embedding strategy which focuses on ambiguous regions by metric learning. The PHRE contains two main processes: ambiguous region pair sampling and hard-aware region embedding, as shown in Figure 4.

Ambiguous Region Pair Sampling. For metric learning, the most important procedure is to sample positive pairs of images which contain the same type of parsing target, and negative pairs of images that have ambiguous targets. We first define an ambiguous graph G with the assistance of an off-the-shelf semantic segmentation method, i.e., PSPNet [40]. Given a human parsing dataset such as LIP [10] or CIHP [9], the PSPNet is trained on the training set to obtain a base model. Then, we utilize the PSPNet to obtain parsing results on the validation set. Based on the results, a normalized confusion matrix is calculated to measure the ambiguity between each pair of classes. We let each parsing class as the node of G and add a directed edge between a pair of classes if their confusion rate is larger than a threshold τ. After that, we discard the isolated nodes with no ambiguous neighbor and obtain the final G, as shown in Figure 4. Given G, we can first sample an image I_a, with class $c_a \in G$, as the anchor, then sample an image I_p with the same class $c_a \in G$ to build a positive pair and an image I_n with a different class $c_b \in G$ to obtain a negative pair. In the next embedding procedure, we consider a positive pair and a negative pair as a triplet $T_I =< I_a, I_p, I_n >$ with the ground truth maps $T_Y =< Y_a, Y_p, Y_n >$.

Hard-aware Region Embedding. During training of the BraidNet, we perform the hard-aware region embedding only for the semantic abstracting net since it can learn semantic knowledge of parsing targets. To measure the region-to-region distance in our embedding method, we need an effective regional feature representation. As shown in Figure 4, we feed an image $I \in T_I$ into the semantic abstracting net to obtain the feature maps of the last convolution layer, denoted as $f(I) \in \mathbb{R}^{w \times h \times k}$, and obtain the predicted probability maps of the softmax layer, denoted as $\hat{Y} \in \mathbb{R}^{w \times h \times C}$, where h and w are the width and height of image I, k is the channel number of the feature map $f(I)$, and C is the number of classes. After that, given the target class c of image I and the predicted probability map $\hat{Y} \in \mathbb{R}^{w \times h \times C}$, we can obtain a hard-aware mask

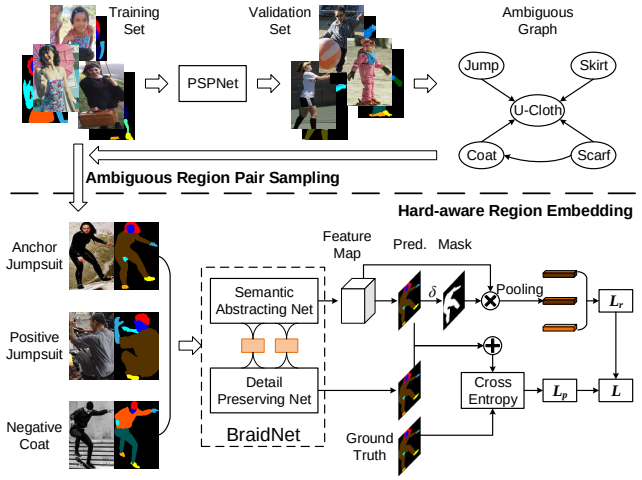

Figure 4: The process of Pairwise Hard Region Embedding.

$M \in \mathbb{R}^{w \times h}$ by:

$$M_{i,j} = \mathbf{1}(\hat{Y}_{i,j,c} < \delta), \tag{1}$$

where δ is the hard-aware threshold, $\mathbf{1}(\cdot)$ returns 1 if the input is true, and 0 otherwise. With the mask M and the feature map $f(I) \in \mathbb{R}^{w \times h \times k}$, we can calculate the regional feature $f_r(I) \in \mathbb{R}^k$ by channel-wise global average pooling as:

$$\begin{aligned} g(f(I)_k, M) &= \frac{\sum_{i=1}^{w} \sum_{j=1}^{h} f(I)_k \odot M}{|M|}, \\ f_r(I) &= [g(f(I)_1, M), g(f(I)_2, M), \ldots, g(f(I)_k, M)], \end{aligned} \tag{2}$$

where $f(I)_k$ is the k-th channel of the feature map f(I) and $\odot$ is the element-wise multiplication. By this means, we can denote the regional features of images in T_I as $T_r =< f_r(I_a), f_r(I_p), f_r(I_n) >$. Next, a hard-aware region embedding loss L_r can be calculated as:

$$\begin{aligned} L_r =& max(0, ||f_r(I_a) - f_r(I_n)|| - ||f_r(I_a) - f_r(I_p)|| - m) \\ &+ \alpha \cdot ||f_r(I_a) - f_r(I_p)||, \end{aligned} \tag{3}$$

where $||\cdot||$ is the L_2 norm, m is a margin to control the constraint, α is a hyper-parameter to balance the constraint on inter-class and intra-class distances.

Moreover, for conventional semantic segmentation and human parsing, the pixel-level cross entropy loss is usually adopted to optimize the networks. Therefore, we also adopt the pixel-to-pixel class label as the basic supervision to compute the pixel loss, which is formulated as:

$$L_p = \sum_{i=1}^{h \times w} \sum_{c=1}^{C} -Y_{i,c} log(\hat{Y}_{i,c}), \tag{4}$$

where H and W are the width and height of the input image, C is the number of classes, $y_{i,c}$ is the ground truth probability of class c at the i-th pixel, and $\hat{y}_{i,c}$ is the predicted probability of class c at the pixel. Finally, we optimize the BraidNet by the combination of hard-aware region embedding loss and the cross entropy loss, which is formulated as:

$$L = L_p + \beta \cdot L_r, \tag{5}$$

where β is a hyper-parameter to balance the two types of losses.

4 EXPERIMENTS

4.1 Datasets and Experimental Setting

In our experiment, we first compare our BraidNet with the state-of-the-art single human parsing methods and conduct the ablation study on the LIP dataset [10]. Then we integrate our method with the instance segmentation method, i.e., Mask R-CNN [11] to evaluate multi-human parsing on the CIHP dataset [9].

The LIP dataset has 50,462 images for single human parsing. Each image contains one person or a part of one person with pixel-level annotation. The images are annotated with 19 semantic human parsing targets and one background class. The dataset is divided into 30,462, 10,000, and 10,000 images for training, validation, and testing, respectively. All compared methods are trained on the training set and evaluated on the validation set, since the testing set is held by the authors for the LIP challenge. We adopt pixel accuracy (Pixel Acc), mean accuracy (Mean Acc), and mean intersection over union (mIoU) for single human parsing following [10].

The CIHP dataset contains 38,280 images for multiple human parsing. Each image contains more than one person. Each person in one image is annotated with not only 20 semantic parts but also a unique instance-level ID. As in [9], we use mIoU to evaluate the performance of region-level parsing The mean average precision at different IoU thresholds (AP^r_{IoU}) and mean of AP^r over IoU $\in [0.1, 0.9]$ with 0.1 interval (AP^r_{vol}) are calculated for instance-level human parsing.

4.2 Implementation Details

This section presents the details on data preparation and training strategy of the networks.

Data Preparation. For single human parsing, the input image for the BraidNet is 384×384. During training, we adopt random scaling, random rotation, horizontal flipping, and random cropping/padding for data augmentation following the strategies. For ambiguous region pair sampling in Section 3.3, the threshold τ in building the ambiguous graph G is set to 0.1 The topology of the graph used to sample image pairs from the LIP dataset is shown in Figure 5.

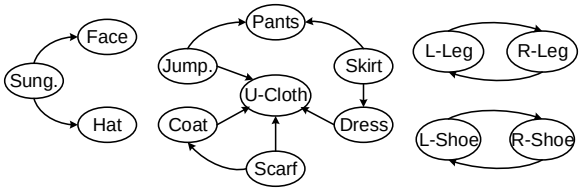

Figure 5: The topology of the ambiguous graph for the LIP dataset.

Networks Training. Before training the whole network of BraidNet, we first train the semantic abstracting net with cross entropy loss by the Stochastic Gradient Descent (SGD) optimizer [16] for 100 epochs. The initial learning rate (LR) is set to 0.01 and adjusted by the "ploy" policy as in [2]. After that, we train the whole BraidNet with the loss function in Equation 5 by the SGD optimizer for 50 epochs. The LR is set to 0.05 and adjusted by the "ploy" policy.

Table 2: The IoU of each class and the mIoU on the LIP dataset.

Method	bg.	hat	hair	glove	sungl.	u-clot.	dress	coat	socks	pants	jumps.	scarf	skirt	face	l-arm	r-arm	l-leg	r-leg	l-shoe	r-shoe	mIoU
DeepLab [2]	84.1	59.8	66.2	28.8	23.9	65.0	33.7	52.9	37.7	68.0	26.1	17.4	25.2	70.0	50.4	53.9	39.4	38.3	27.0	28.4	44.8
PSPNet [40]	86.1	63.5	68.0	39.1	23.8	68.1	31.7	56.2	44.5	72.7	28.7	15.7	25.7	70.8	59.7	62.3	54.9	54.5	42.3	42.9	50.6
MMAN [29]	84.8	57.7	65.6	30.1	20.0	64.2	28.4	52.0	41.5	71.0	23.6	9.7	23.2	69.5	55.3	58.1	51.9	52.2	38.6	39.0	46.8
JPPNet [17]	86.3	63.6	70.2	36.2	23.5	68.2	31.4	55.7	44.6	72.2	28.4	18.8	25.1	73.4	62.0	63.9	58.2	58.0	44.0	44.1	51.4
CE2P [22]	87.4	64.6	**72.1**	38.4	**32.2**	68.9	32.2	55.6	48.8	73.5	27.2	13.8	22.7	74.9	64.0	65.9	59.7	58.0	45.7	45.6	52.6
BraidNet (ours)	**88.0**	**66.8**	72.0	**42.5**	32.1	**69.8**	**33.7**	**57.4**	**49.0**	**74.9**	**32.4**	**19.3**	**27.2**	**74.9**	**65.5**	**67.9**	**60.2**	**59.6**	**47.4**	**47.9**	**54.4**

4.3 Single Human Parsing

To validate the effectiveness of the proposed BraidNet with PHRE strategy, we compare it with several state-of-the-art methods on the LIP dataset. The details of methods are as follows:

1) Pyramid Scene Pooling Network (PSPNet) [40]. The PSPNet is one of the state-of-the-art frameworks for semantic segmentation. It represents the FCNs that adopt multi-scale feature pooling in one single network. We implement the PSPNet with ResNet-101 [12] as the backbone for single human parsing.

2) DeepLab [3]. The DeepLab is also one of the state-of-the-art semantic segmentation method. It adopts the atrous convolution with different dilation rates to capture multi-scale features. DeepLab also uses the conditional random field to refine segmentation results. We directly use the parsing results of DeepLab on LIP in [17].

3) Joint Parsing & Pose Estimation Network (JPPNet) [17]. JPPNet has a multi-task learning framework for both human parsing and pose estimation. It adopts ResNet-101 [12] as the backbone of the network. We also refer to the results on the LIP dataset in [17].

4) Macro-Micro Adversarial Network (MMAN) [29]. MMAN is a generative model for human parsing. It adopts the GAN-based framework which has one generator using the DeepLab as the backbone to output parsing results and two discriminators to concentrate on macro features and micro details, respectively.

5) Context Embedding and Edge Preserving (CE2P) [22]. CE2P is the state-of-the-art human parsing approach on the LIP dataset. This method utilizes the PSPNet as the basic model. The authors adopts a context embedding branch to combine low-level features with the context features. CE2P also has an edge preserving branch to make the results have sharp edges and details.

6) CE2P (w flip) [22]. Different from other methods, this is the CE2P while applying horizontal flipping on the input image during testing. The final result is the combination of results from the original input and the flipped image. We refer to the results from [22].

7) Braiding Network (BraidNet). This is our proposed Braiding Network with Pairwise Hard Region Embedding.

The Pixel Acc, Mean Acc, and mIoU of these methods are listed in Table 3. We can first find that the general method for semantic segmentation, i.e., DeepLab and PSPNet, are relatively worse than human parsing methods. Because these method does not consider specific characteristics of human parsing task. They can be adopted as the baseline or basic model for human parsing method. Since MMAN utilizes DeepLab as the backbone to build a GAN for parsing mask generation, it obtains the better result than DeepLab. However, it is still worse than other methods since the GAN framework may be difficult to be optimized for dense prediction task. Moreover, the JPPNet achieves the better result than general semantic segmentation methods with a large margin. This proves that the pose estimation task can effectively improve human parsing by joint learning. The CE2P has the better results than above methods, as it adopts a PSPNet-based FCN with a context embedding branch to combine multi-level features and an edge preserving branch to capture more details near to the edge. Finally, our BraidNet outperforms the state-of-the-art methods, even the CE2P with flipping. The results demonstrate that the BraidNet can learn robust and discriminative representation for human parsing.

Table 3: Comparison of the state-of-the-art single human parsing methods on the LIP dataset.

Method	Pixel Acc	Mean Acc	mIoU
DeepLab [2]	84.09	55.62	44.80
PSPNet [40]	86.23	61.33	50.56
MMAN [29]	-	-	46.81
JPPNet [17]	86.39	62.32	51.37
CE2P [22]	-	-	52.56
CE2P (w flip) [22]	87.37	63.20	53.10
BraidNet (ours)	**87.60**	**66.09**	**54.42**

Table 2 lists the comparison of per-class and mean IoU results by which we may find more interesting phenomenons. First of all, for all methods, large parsing targets such as hat, hair, upper-clothes, pants, face, and arms are easier for parsing. One reason is that large targets usually have more pixel samples for training a deep network. The other is that large targets may overwhelm small objects in forward propagation of convolution networks. The second finding is that the classes with scarce samples obtain poor results, even for large objects like dress, jumpsuit, and skirt. At last, we can find that our method achieves significant improvement for several difficult classes, e.g., glove, dress, jumpsuit, and skirt. This demonstrates that our Pairwise Hard Region Embedding can make the BraidNet learn more discriminative features for these ambiguous parsing targets.

4.4 Ablation Study on BraidNet

In this section, we provide analysis on the margin m in Equation 3 during training the modules in the BraidNet.

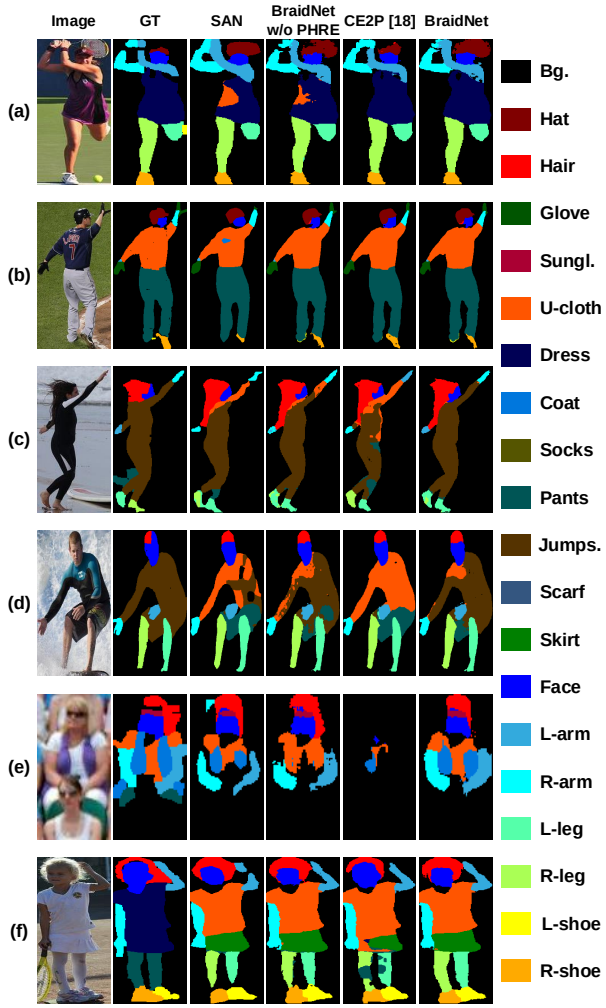

Figure 6: Visualization of parsing results on the LIP dataset.

We first provide the analysis on the margin in PHRE, as listed in Table 4. As discussed in Section 3.3, the margin m is a constraint between inter-class distance and intra-class distance. From the results, we find that $m = 2.0$ is suitable for our task. Small margin cannot push the samples from different classes to separate. While too big margin will make the network difficult to optimize. Therefore, we select $m = 2.0$ in our implementation.

We then conduct the ablation study on the BraidNet. The combinations of the semantic abstracting net, the detain preserving net, the braiding Module, and the PHRE are as follows:

1) Semantic Abstracting Net (SAN). This is the SAN with the modified PSPNet as the backbone which is trained by the cross entropy loss.

2) detail-preserving Net (DPN). This is the DPN trained by the cross entropy loss.

3) SAN + DPN. The outputs of SAN and DPN are directly fused with the weight of 0.5 : 0.5.

4) BraidNet without PHRE (BraidNet w/o PHRE). This is our BraidNet which adopts the braiding module to connect SAN and DPN. We use the cross entropy loss to train this framework.

Table 4: Analysis on the margin in PHRE.

Margin	Pixel Acc	Mean Acc	mIoU
1.0	87.56	65.55	54.06
2.0	**87.60**	**66.09**	**54.42**
3.0	87.50	65.64	54.00
4.0	87.52	65.31	53.91
5.0	87.49	65.42	53.79

Table 5: The ablation study of BraidNet on the LIP dataset.

Method	Pixel Acc	Mean Acc	mIoU
SAN	86.23	61.33	50.56
DPN	59.81	10.27	7.10
SAN + DPN	85.52	53.62	46.96
SAN w PHRE	86.36	63.42	51.73
BraidNet w/o PHRE	87.44	64.13	53.19
BraidNet	**87.60**	**66.09**	**54.42**

5) SAN with PHRE (SAN w PHRE). In this scheme, we train the SAN with the pairwise hard region embedding.

6) BraidNet. This is the whole framework of Braiding Network with Pairwise Hard Region Embedding.

The results of the ablation study are listed in Table 5. From the comparison of SAN and DPN, we can find that DPN has very poor results. Because the DPN is focused on local details, while SAN concentrates on high-level semantics. This shows that semantic information is more important than texture for human parsing, since it is a pixel-level classification task which needs more semantics. The BraidNet without PHRE is better than late fusion of SAN and DPN. This means that the direct combination of the outputs can hardly effectively exploit the semantics learned by SAN and the details from DPN. While the braiding module in the BraidNet provides a bridge to exchange complementary information between the two sub-nets. At last, by comparing SAN and BraidNet w/o PHRE with the SAN w PHRE and BraidNet, we can see that the proposed PHRE can play an important role to improve the parsing results. The ablation study can demonstrate the effectiveness of each module in our BraidNet.

We also show some parsing results of SAN, BraidNet without PHRE, CE2P [22], and the BraidNet in Figure 6. Overall, our BraidNet and CE2P achieve competitive results under varied human poses and viewpoints, as shown in Figure 6(a) and (b). While the BraidNet performs better for details and some ambiguous target. For instance, in Figure 6(c) and (d), the BraidNet can effectively distinguish jumpsuits from upper-clothes and accurately generate the edges of hands, gloves, hat, and hair. Moreover, our BraidNet obtains excellent parsing result on low-resolution images like Figure 6(e), while the CE2P almost fails on this condition. In addition, all methods have the capacity of generalization, as shown in Figure 6(f). Although there are some noisy images with mistakes of annotation in LIP, the models can also output correct results.

By comparing SAN with the two BraidNet, we can see that the results of SAN have more smooth edges and miss some small targets, while the results of BraidNet have more details. This not only demonstrates the effectiveness of the detail-preserving net which

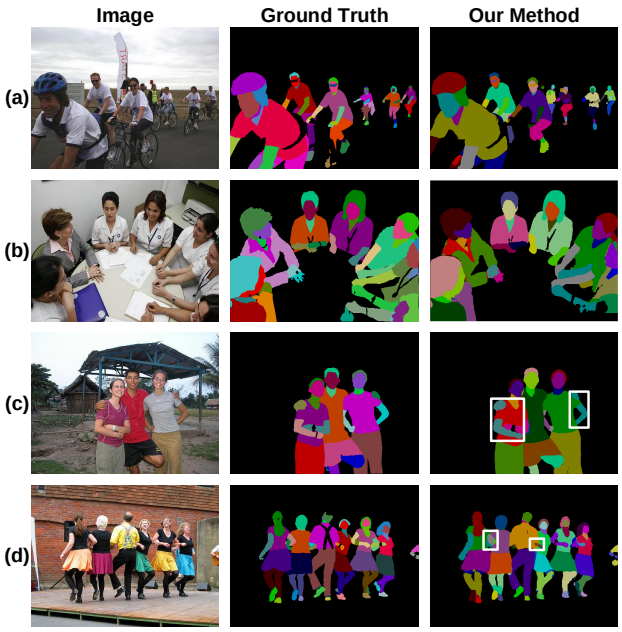

Figure 7: Examples of results from the BraidNet + Mask R-CNN on the CIHP testing dataset.

can capture more details but also validates that the braiding module can comprehensively exploit the semantics and the details from the two networks of the BraidNet. Finally, the PHRE strategy further improve the performance of the BraidNet, which can differentiate the similar parsing targets.

4.5 Multiple Human Parsing

Our BraidNet can be seamlessly integrated with the off-the-shelf instance segmentation framework to perform instance-level multiple human parsing. In this section, we directly adopt a Mask R-CNN [11] trained on MS-COCO dataset [19] to segment human instances from images. For each segmented instance, we utilize the BraidNet to obtain the human parsing result. Moreover, we also take the whole image as the input for the BraidNet to obtain a global parsing result. At last, the instance-level and global results are combined by late fusion. We compare our results with several state-of-the-art methods on the CIHP testing set. The details of methods are as follows:

1) Part Grouping Network (PGN) [9]. The PGN represents the detection-free framework for multiple human parsing. It contains a multi-task deep FCN to output the instance-diagnostic parsing result and the contours of the human instances simultaneously. Finally, an instance partition process is employed to obtain the final instance-level parsing results.

2) M-CE2P [22]. The M-CE2P is the state-of-the-art multi-human parsing method, which integrates the CE2P with Mask R-CNN [11]. For fairness, we use adopt same Mask R-CNN model as our method to segment instance for M-CE2P. Then, we use the global parsing and two local parsing models released by the authors of [22] for human parsing. The results of the three models are aggregated and refined to obtain the final results as in [22].

Table 6: Comparison of the state-of-the-art multiple human parsing methods on the CIHP set.

Method	mIoU	$AP^r_{0.5}$	$AP^r_{0.6}$	$AP^r_{0.7}$	AP^r_m
PGN [9]	55.80	35.80	28.60	20.50	33.60
M-CE2P [22]	59.50	48.69	40.13	29.74	42.83
BraidNet+Mask R-CNN	**60.62**	**48.99**	**41.67**	**32.71**	**43.59**

3) BraidNet + Mask R-CNN. This is our framework for multiple human parsing.

The results of the methods are listed in Table 6. From the results, we can see that our method and M-CE2P achieves much better results than PGN. This proves that the top-down framework, i.e., first detecting and segmenting person instance then parsing human parts, is more effective than the detection-free method. Moreover, our framework outperforms M-CE2P and achieves the state-of-the-art results on the CIHP dataset. In particular, the BraidNet has better performance under large IoU thresholds, which demonstrates the effectiveness of our approach. At last, we also show several parsing results of our method on the CIHP dataset, as shown in Figure 7. We can find that our method also obtains excellent results for multiple human parsing, especially for the small instances. However, there are also some failure cases, as shown in Figure 7 (c) and (d). The main reason is the occlusion and intersection between body parts of different persons.

In the future work, we will further explore how to effectively aggregate the results of instance-level segmentation and the part-level parsing by considering the global and local relations among human parts [26, 33]. Furthermore, our method may also be applied to video human parsing by adopting efficient models with temporal information for video analysis [20, 42].

5 CONCLUSION

In this paper, we propose a Braiding Network, named as BraidNet, with Pairwise Hard Region Embedding strategy for fine-grained human parsing. To learn discriminative representation, the BraidNet has two parallel sub-nets to model semantic knowledge and local structures, respectively. Specifically, the semantic abstracting net contains a narrow-down architecture to learn high-level semantics from pixels. While the detail-preserving net has a wide but shallow structure without down-sampling to capture more detailed texture for small objects. Moreover, an elaborated braiding module is inserted between the two sub-nets to make them exchange complementary information during training. Furthermore, we propose a pairwise PHRE strategy which can discover the ambiguous parsing targets with regional embedding. Therefore, our BraidNet can effectively exploit the multi-level information to learn a robust representation for fine-grained human parsing. In addition, the BraidNet can be seamlessly integrated with instance segmentation method for the instance-level multi-human parsing task. Extensive experiments on the public datasets demonstrate the effectiveness of the proposed framework.

6 ACKNOWLEDGEMENT

This work is partially supported by the National Science Foundation of China (No. 61602049).

REFERENCES

[1] Vijay Badrinarayanan, Alex Kendall, and Roberto Cipolla. 2017. SegNet: A Deep Convolutional Encoder-Decoder Architecture for Image Segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* 39, 12 (2017), 2481–2495.

[2] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L. Yuille. 2018. DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs. *IEEE Trans. Pattern Anal. Mach. Intell.* 40, 4 (2018), 834–848.

[3] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. 2018. Encoder-Decoder with Atrous Separable Convolution for Semantic Image Segmentation. In *European Conference on Computer Vision*. 833–851.

[4] Jifeng Dai, Haozhi Qi, Yuwen Xiong, Yi Li, Guodong Zhang, Han Hu, and Yichen Wei. 2017. Deformable Convolutional Networks. In *IEEE International Conference on Computer Vision*. 764–773.

[5] Haoshu Fang, Guansong Lu, Xiaolin Fang, Jianwen Xie, Yu-Wing Tai, and Cewu Lu. 2018. Weakly and Semi Supervised Human Body Part Parsing via Pose-Guided Knowledge Transfer. In *IEEE Conference on Computer Vision and Pattern Recognition*. 70–78.

[6] Chuang Gan, Zhe Gan, Xiaodong He, Jianfeng Gao, and Li Deng. 2017. Stylenet: Generating attractive visual captions with styles. In *CVPR*. 955–964.

[7] Chuang Gan, Ming Lin, Yi Yang, Gerard De Melo, and Alexander G Hauptmann. 2016. Concepts not alone: Exploring pairwise relationships for zero-shot video activity recognition. In *AAAI Conference on Artificial Intelligence*. 3487–3493.

[8] Chuang Gan, Naiyan Wang, Yi Yang, Dit-Yan Yeung, and Alex G Hauptmann. 2015. Devnet: A deep event network for multimedia event detection and evidence recounting. In *IEEE International Conference on Computer Vision and Pattern Recognition*. 2568–2577.

[9] Ke Gong, Xiaodan Liang, Yicheng Li, Yimin Chen, Ming Yang, and Liang Lin. 2018. Instance-Level Human Parsing via Part Grouping Network. In *European Conference Computer Vision*. 805–822.

[10] Ke Gong, Xiaodan Liang, Dongyu Zhang, Xiaohui Shen, and Liang Lin. 2017. Look into Person: Self-Supervised Structure-Sensitive Learning and a New Benchmark for Human Parsing. In *IEEE Conference on Computer Vision and Pattern Recognition*. 6757–6765.

[11] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross B. Girshick. 2017. Mask R-CNN. In *IEEE International Conference on Computer Vision*. 2980–2988.

[12] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep Residual Learning for Image Recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*. 770–778.

[13] Alexander Hermans, Lucas Beyer, and Bastian Leibe. 2017. In Defense of the Triplet Loss for Person Re-Identification. *CoRR* abs/1703.07737 (2017). http://arxiv.org/abs/1703.07737

[14] Mahdi M. Kalayeh, Emrah Basaran, Muhittin Gökmen, Mustafa E. Kamasak, and Mubarak Shah. 2018. Human Semantic Parsing for Person Re-Identification. In *IEEE Conference on Computer Vision and Pattern Recognition*. 1062–1071.

[15] Simon Kohl, Bernardino Romera-Paredes, Clemens Meyer, Jeffrey De Fauw, Joseph R. Ledsam, Klaus H. Maier-Hein, S. M. Ali Eslami, Danilo Jimenez Rezende, and Olaf Ronneberger. 2018. A Probabilistic U-Net for Segmentation of Ambiguous Images. In *Advances in Neural Information Processing Systems*. 6965–6975.

[16] Yann LeCun, Léon Bottou, Yoshua Bengio, Patrick Haffner, et al. 1998. Gradient-based learning applied to document recognition. *Proc. IEEE* 86, 11 (1998), 2278–2324.

[17] Xiaodan Liang, Ke Gong, Xiaohui Shen, and Liang Lin. 2019. Look into Person: Joint Body Parsing & Pose Estimation Network and a New Benchmark. *IEEE Trans. Pattern Anal. Mach. Intell.* 41, 4 (2019), 871–885.

[18] Xiaodan Liang, Chunyan Xu, Xiaohui Shen, Jianchao Yang, Si Liu, Jinhui Tang, Liang Lin, and Shuicheng Yan. 2015. Human Parsing with Contextualized Convolutional Neural Network. In *IEEE International Conference on Computer Vision*. 1386–1394.

[19] Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. 2014. Microsoft COCO: Common Objects in Context. In *European Conference on Computer Vision*. 740–755.

[20] Kun Liu, Wu Liu, Chuang Gan, Mingkui Tan, and Huadong Ma. 2018. T-C3D: Temporal Convolutional 3D Network for Real-time Action Recognition. In *AAAI Conference on Artificial Intelligence*. 7138–7145.

[21] Si Liu, Xiaodan Liang, Luoqi Liu, Ke Lu, Liang Lin, and Shuicheng Yan. 2014. Fashion Parsing with Video Context. In *ACM International Conference on Multimedia*. 467–476.

[22] Ting Liu, Tao Ruan, Zilong Huang, Yunchao Wei, Shikui Wei, Yao Zhao, and Thomas Huang. 2019. Devil in the Details: Towards Accurate Single and Multiple Human Parsing. In *AAAI Conference on Artificial Intelligence*. 4814–4821.

[23] Wu Liu, Xinchen Liu, Huadong Ma, and Peng Cheng. 2017. Beyond Human-level License Plate Super-resolution with Progressive Vehicle Search and Domain Priori GAN. In *ACM International Conference on Multimedia*. 1618–1626.

[24] Xinchen Liu, Wu Liu, Tao Mei, and Huadong Ma. 2016. A Deep Learning-Based Approach to Progressive Vehicle Re-identification for Urban Surveillance. In *European Conference on Computer Vision*. 869–884.

[25] Xinchen Liu, Wu Liu, Tao Mei, and Huadong Ma. 2018. PROVID: Progressive and Multimodal Vehicle Reidentification for Large-Scale Urban Surveillance. *IEEE Trans. Multimedia* 20, 3 (2018), 645–658.

[26] Xinchen Liu, Wu Liu, Meng Zhang, Jingwen Chen, Lianli Gao, Chenggang Yan, and Tao Mei. 2019. Social Relation Recognition from Videos via Multi-scale Spatial-Temporal Reasoning. In *IEEE International Conference on Computer Vision and Pattern Recognition*.

[27] Jonathan Long, Evan Shelhamer, and Trevor Darrell. 2015. Fully convolutional networks for semantic segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition*. 3431–3440.

[28] Xianghui Luo, Zhuo Su, Jiaming Guo, Gengwei Zhang, and Xiangjian He. 2018. Trusted Guidance Pyramid Network for Human Parsing. In *ACM International Conference on Multimedia*. 654–662.

[29] Yawei Luo, Zhedong Zheng, Liang Zheng, Tao Guan, Junqing Yu, and Yi Yang. 2018. Macro-Micro Adversarial Network for Human Parsing. In *European Conference on Computer Vision*. 424–440.

[30] Sangwoo Mo, Minsu Cho, and Jinwoo Shin. 2019. InstaGAN: Instance-aware Image-to-Image Translation. In *International Conference on Learning Representations*.

[31] Hyeonwoo Noh, Seunghoon Hong, and Bohyung Han. 2015. Learning Deconvolution Network for Semantic Segmentation. In *IEEE International Conference on Computer Vision*. 1520–1528.

[32] Rui Qian, Yunchao Wei, Honghui Shi, Jiachen Li, Jiaying Liu, and Thomas S. Huang. 2019. Weakly Supervised Scene Parsing with Point-based Distance Metric Learning. In *AAAI Conference on Artificial Intelligence*. 8843–8850.

[33] Weijian Ruan, Jun Chen, Yi Wu, Jinqiao Wang, Chao Liang, Ruimin Hu, and Junjun Jiang. 2018. Multi-correlation filters with triangle-structure constraints for object tracking. *IEEE Trans. on Multimedia* 21, 5 (2018), 1122–1134.

[34] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael S. Bernstein, Alexander C. Berg, and Fei-Fei Li. 2015. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision* 115, 3 (2015), 211–252.

[35] Florian Schroff, Dmitry Kalenichenko, and James Philbin. 2015. FaceNet: A unified embedding for face recognition and clustering. In *IEEE Conference on Computer Vision and Pattern Recognition*. 815–823.

[36] Kota Yamaguchi, M. Hadi Kiapour, and Tamara L. Berg. 2013. Paper Doll Parsing: Retrieving Similar Styles to Parse Clothing Items. In *IEEE International Conference on Computer Vision*. 3519–3526.

[37] Kota Yamaguchi, M. Hadi Kiapour, Luis E. Ortiz, and Tamara L. Berg. 2012. Parsing clothing in fashion photographs. In *IEEE Conference on Computer Vision and Pattern Recognition*. 3570–3577.

[38] Fisher Yu and Vladlen Koltun. 2016. Multi-Scale Context Aggregation by Dilated Convolutions. In *International Conference on Learning Representations*.

[39] Yuhui Yuan, Kuiyuan Yang, and Chao Zhang. 2017. Hard-Aware Deeply Cascaded Embedding. In *IEEE International Conference on Computer Vision*. 814–823.

[40] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. 2017. Pyramid Scene Parsing Network. In *IEEE Conference on Computer Vision and Pattern Recognition*. 6230–6239.

[41] Jian Zhao, Jianshu Li, Yu Cheng, Terence Sim, Shuicheng Yan, and Jiashi Feng. 2018. Understanding Humans in Crowded Scenes: Deep Nested Adversarial Learning and A New Benchmark for Multi-Human Parsing. In *ACM International Conference on Multimedia*. 792–800.

[42] Qixian Zhou, Xiaodan Liang, Ke Gong, and Liang Lin. 2018. Adaptive Temporal Encoding Network for Video Instance-level Human Parsing. In *ACM International Conference on Multimedia*. 1527–1535.