

Look into Person: Self-supervised Structure-sensitive Learning and A New Benchmark for Human Parsing

Ke Gong¹, Xiaodan Liang^{1,2}, Dongyu Zhang^{1*}, Xiaohui Shen³, Liang Lin^{1,4}

¹Sun Yat-sen University ²Carnegie Mellon University ³Adobe Research ⁴SenseTime Group (Limited)

gongk3@mail2.sysu.edu.cn, xiaodan1@cs.cmu.edu, {zhangdy27, linlmg}@mail.sysu.edu.cn, xshen@adobe.com

Abstract

Human parsing has recently attracted a lot of research interests due to its huge application potentials. However existing datasets have limited number of images and annotations, and lack the variety of human appearances and the coverage of challenging cases in unconstrained environment. In this paper, we introduce a new benchmark¹ “Look into Person (LIP)” that makes a significant advance in terms of scalability, diversity and difficulty, a contribution that we feel is crucial for future developments in human-centric analysis. This comprehensive dataset contains over 50,000 elaborately annotated images with 19 semantic part labels, which are captured from a wider range of viewpoints, occlusions and background complexity. Given these rich annotations we perform detailed analyses of the leading human parsing approaches, gaining insights into the success and failures of these methods. Furthermore, in contrast to the existing efforts on improving the feature discriminative capability, we solve human parsing by exploring a novel self-supervised structure-sensitive learning approach, which imposes human pose structures into parsing results without resorting to extra supervision (i.e., no need for specifically labeling human joints in model training). Our self-supervised learning framework can be injected into any advanced neural networks to help incorporate rich high-level knowledge regarding human joints from a global perspective and improve the parsing results. Extensive evaluations on our LIP and the public PASCAL-Person-Part dataset demonstrate the superiority of our method.

1. Introduction

Human parsing aims to segment a human image into multiple parts with fine-grained semantics and provide more detailed understanding of image contents. It can stimulate

many higher-level computer vision applications [33], such as person re-identification [34] and human behavior analysis [11, 16].

Recently, Convolutional Neural Networks (CNNs) have achieved exciting success in human parsing [13, 17, 15]. Nevertheless, as demonstrated in many other problems such as object detection [14] and semantic segmentation [35], the performance of those CNN-based approaches heavily rely on the availability of annotated images for training. In order to train a human parsing network with potentially practical value in real-world applications, it is highly desired to have a large-scale dataset composed of representative instances with varied clothing appearances, strong articulation, partial (self-)occlusions, truncation at image borders, diverse viewpoints and background clutters. Although there exist training sets for special scenarios such as fashion pictures [30, 9, 13, 17] and people in constrained situations (e.g., upright) [6], these datasets are limited in their coverage and scalability, as shown in Fig. 1. The largest public human parsing dataset [17] so far only contains 17,000 fashion images while others only include thousands of images.

Moreover, to the best of our knowledge, no attempt has been made to establish a standard representative benchmark aiming to cover a wide pallet of challenges for the human parsing task. The existing datasets did not provide an evaluation server with a secret test set to avoid potential dataset over-fitting, which hinders further development on this topic. Therefore we propose a new benchmark “Look into Person (LIP)” and a public server for automatically reporting evaluation results. Our benchmark significantly advances the state-of-the-arts in terms of appearance variability and complexity, which includes 50,462 human images with pixel-wise annotations of 19 semantic parts.

The recent progress on human parsing [5, 28, 31, 30, 9, 25, 19, 17] has been achieved by improving the feature representations using convolutional neural networks and recurrent neural networks. To capture rich structure information, they combine CNNs and the graphical models (e.g., Conditional Random Fields (CRFs)), similar to the general object segmentation approaches [35, 4, 27]. However, evaluated

*The first two authors contribute equally to this paper. Corresponding author is Dongyu Zhang. This work was supported by the National Natural Science Foundation of China under Grant 61401125 and 61671182.

¹The dataset is available at <http://hcp.sysu.edu.cn/lip>

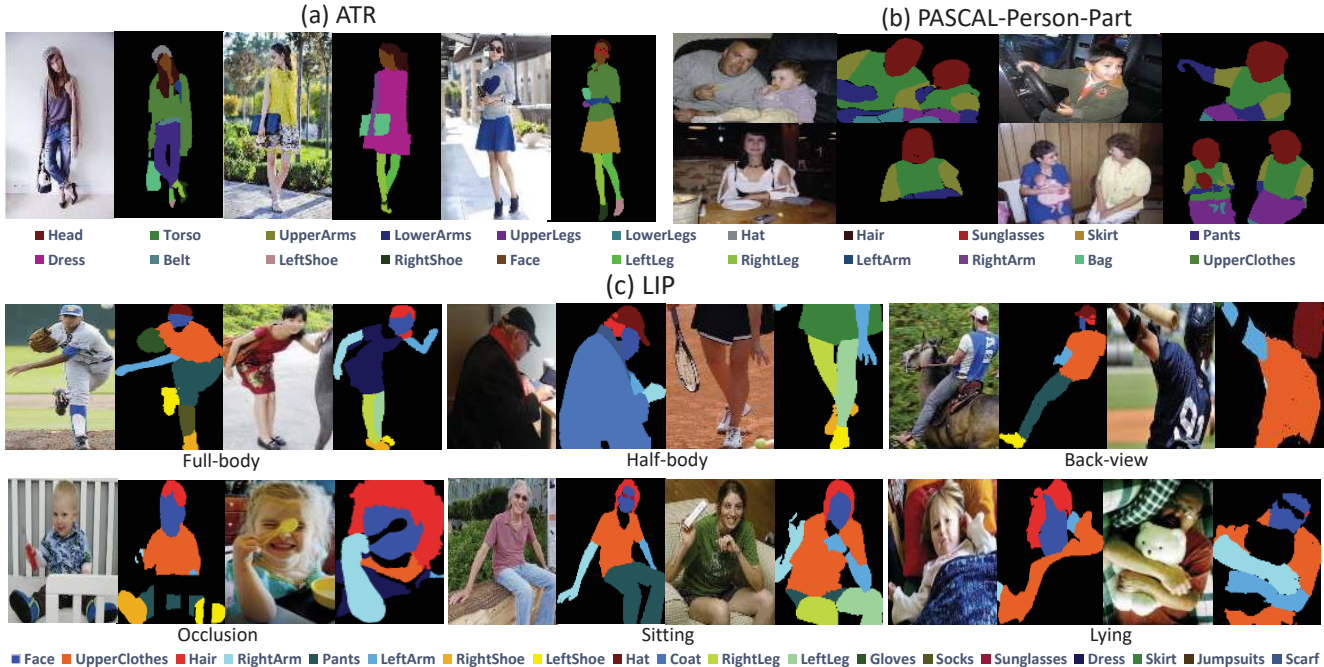


Figure 1: Annotation examples for our “Look into Person (LIP)” dataset and existing datasets. (a) The images in ATR dataset which are fixed in size and only contain stand-up person instances in the outdoors. (b) The images in PASCAL-Person-Part dataset which also have lower scalability and only contain 6 coarse labels. (c) The images in our LIP dataset with high appearance variability and complexity.

on the new LIP dataset, the results of some existing methods [3, 21, 4, 5] are unsatisfactory. Without imposing human body structure priors, these general approaches based on bottom-up appearance information sometimes tend to produce unreasonable results (e.g., right arm connected with left shoulder), as shown in Fig. 2. The human body structural information has been previously well-explored in the human pose estimation [32, 7] where dense joint annotations are provided. However, since human parsing requires more extensive and detailed prediction than pose estimation, it is difficult to directly utilize joint-based pose estimation models in pixel-wise prediction to incorporate the complex structure constraints. In order to explicitly enforce the produced parsing results to be semantically consistent with the human pose / joint structures, we propose a novel structure-sensitive learning approach for human parsing. In addition to using the traditional pixel-wise part annotations as the supervision, we introduce a structure-sensitive loss to evaluate the quality of predicted parsing results from a joint structure perspective. That means a satisfactory parsing result should be able to preserve a reasonable joint structure (e.g., the spatial layouts of human parts). Note that annotating both pixel-wise labeling map and pose joints is expensive and may cause ambiguities. Therefore in this work we generate approximated human joints directly from the parsing annotations and use them as the supervision signal for the structure-sensitive loss, which is hence called a “self-supervised” strategy, noted as Self-supervised Structure-

| Dataset | #Training | #Validation | #Test | Categories |
|------------------------|---------------|---------------|---------------|------------|
| Fashionista [31] | 456 | - | 229 | 56 |
| PASCAL-Person-Part [6] | 1,716 | - | 1,817 | 7 |
| ATR [17] | 16,000 | 700 | 1,000 | 18 |
| LIP | 30,462 | 10,000 | 10,000 | 20 |

Table 1: Overview of the publicly available datasets for human parsing. For each dataset we report the number of annotated persons in training, validation and test sets as well as the number of categories including background.

sensitive Learning (SSL).

Our contributions are summarized in the following three aspects. 1) We propose a new large-scale benchmark and an evaluation server to advance the human parsing research, in which 50,462 images with pixel-wise annotations on 19 semantic part labels are provided. 2) By experimenting on our benchmark, we present the detailed analyses about the existing human parsing approaches to gain some insights into the success and failures of these approaches. 3) We propose a novel self-supervised structure-sensitive learning framework for human parsing, which is capable of explicitly enforcing the consistency between the parsing results and the human joint structures. Our proposed framework significantly surpasses the previous methods on both the existing PASCAL-Person-Part dataset [6] and our new LIP dataset.

1.1. Related Work

Human parsing datasets: The commonly used publicly available datasets for human parsing are summarized in Table. 1. The previous datasets were labeled with limited

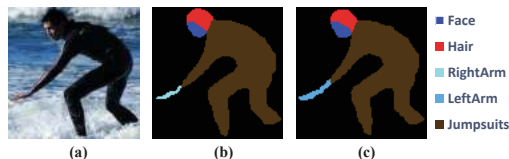


Figure 2: An example shows that self-supervised structure-sensitive learning is helpful for human parsing. (a): The original image. (b): The parsing results by Attention-to-scale [5] where the left-arm is wrongly labeled as right-arm. (c): Our parsing results successfully incorporate the structure information to generate reasonable outputs.

number of images or categories. Containing 50,462 images annotated with 20 categories, our LIP dataset is the largest and most comprehensive human parsing dataset to date. Some other datasets in the vision community were dedicated to the tasks of clothes recognition, retrieval [20, 23] and human pose estimation [1, 12], while our LIP dataset only focuses on human parsing.

Human parsing approaches: Recently many research efforts have been devoted to human parsing [17, 31, 30, 25, 19, 28, 5]. For example, Liang *et al.* [17] proposed a novel Co-CNN architecture which integrates multiple levels of image contexts into a unified network. Besides human parsing, there has also been increasing research interest on the part segmentation of other objects such as animals or cars [26, 27, 22]. To capture the rich structure information based on the advanced CNN architecture, common solutions include the combination of CNNs and CRFs [4, 35] and the adoptions of multi-scale feature representations [4, 5, 28]. Chen *et al.* [5] proposed an attention mechanism that learns to weight the multi-scale features at each pixel location. Some previous works [8, 29] explored human pose information to guide human parsing by generating “pose-guided” part segment proposals. To leverage human joint structure more effortlessly and efficiently, the focus in our approach is nevertheless a new self-supervised structure-sensitive learning approach, which actually can be embedded in any networks.

2. Look into Person Benchmark

In this section we introduce our new “Look into Person (LIP)”, a new large-scale dataset focusing on semantic understanding of human bodies which has several appealing properties. First, with 50,462 annotated images, LIP is an order of magnitude larger and more challenging than previous similar attempts [31, 6, 17]. Second, LIP is annotated with elaborated pixel-wise annotations with 19 semantic human part labels and one background label. Third, the images collected from the real-world scenarios contain people appearing with challenging poses and viewpoints, heavy occlusions, various appearances and in wide range of resolutions. Furthermore, the background of images in the LIP dataset is also more complex and diverse than the one in

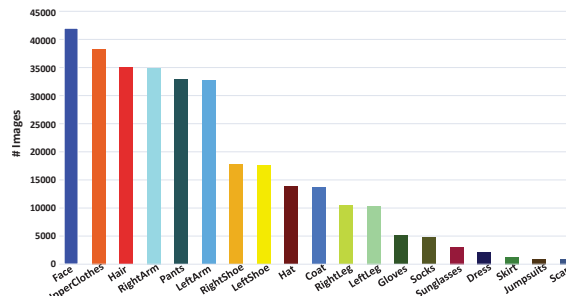


Figure 3: The data distribution on 19 semantic part labels in the LIP dataset.

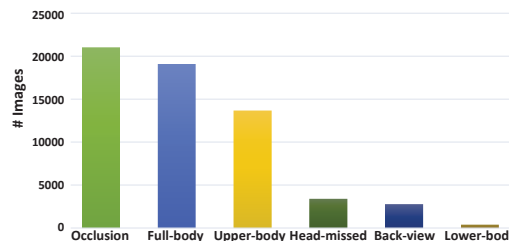


Figure 4: The numbers of images that show diverse visibilities in the LIP dataset, including occlusion, full-body, upper-body, lower-body, head-missed and back-view.

previous datasets. Some examples are showed in Fig. 1. With the LIP dataset, we propose a new benchmark suite for human parsing together with a standard evaluation server where the test set will be kept secret to avoid overfitting.

2.1. Image Annotation

The images in the LIP dataset are cropped person instances from Microsoft COCO [18] training and validation sets. We defined 19 human parts or clothes labels for annotation, which are hat, hair, sunglasses, upper-clothes, dress, coat, socks, pants, gloves, scarf, skirt, jumpsuits, face, right arm, left arm, right leg, left leg, right shoe, left shoe, and in addition to a background label. We implemented an annotation tool and generated multi-scale superpixels of images based on [2] to speed up the annotation.

2.2. Dataset splits

In total, there are 50,462 images in the LIP dataset including 19,081 full-body images, 13,672 upper-body images, 403 lower-body images, 3,386 head-missed images, 2,778 back-view images and 21,028 images with occlusions. We split the images into separate training, validation and test sets. Following random selection, we arrive at a unique split consisting of 30,462 training and 10,000 validation images with publicly available annotations, as well as 10,000 test images with annotations withheld for benchmarking purpose.

2.3. Dataset statistics

In this section we analyse the images and categories in the LIP dataset in detail. In general, face, arms and legs

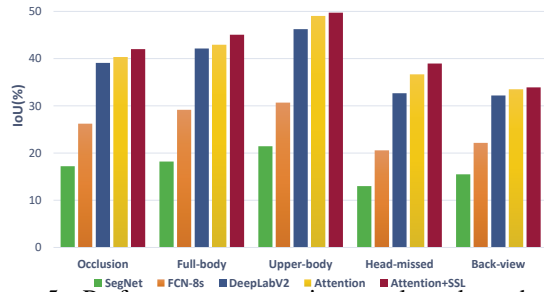


Figure 5: Performance comparison evaluated on the LIP validation set with different appearance, including occlusion, full-body, upper-body, head-missed and back-view.

are the most remarkable parts of a human body. However, human parsing aims to analyse every detailed regions of a person including different body parts as well as different categories of clothes. We therefore define 6 body parts and 13 clothes categories. Among these 6 body parts, we divide arms and legs into left side and right side for more precise analysis, which also increases the difficulty of the task. As for clothes classes, we have not only common clothes like upper clothes, pants and shoes, but also infrequent categories such as skirts and jumpsuits. Furthermore, small scale accessories like sunglasses, gloves and socks are also taken into account. The numbers of images for each semantic part label are presented in Fig. 3

The images in the LIP dataset contain diverse human appearances, viewpoints and occlusions. Additionally, more than half of the images suffer occlusions of different degrees. Occlusion is considered occurred if any of the 19 semantic parts appears in the image but is occluded or invisible. In more challenging cases, the images contain person instances in a back-view, which gives rise to more ambiguity of left and right spatial layouts. The numbers of images of different appearance (i.e. occlusion, full-body, upper-body, head-missed, back-view and lower-body) are summarized in Fig. 4.

3. Empirical study of state-of-the-arts

In this section we analyse the performance of leading human parsing or semantic object segmentation approaches on our benchmark. We take advantage of our rich annotations and conduct a detailed analysis of various factors influencing the results, such as appearance, foreshortening and viewpoints. The goal of this analysis is to evaluate the robustness of the current approaches in various challenges for human parsing, and identify the existing limitations to stimulate further research advances.

In our analysis, we consider fully convolutional networks [21] (FCN-8s), a deep convolutional encoder-decoder architecture [3] (SegNet), deep convolutional nets with atrous convolution and multi-scale [4] (DeepLabV2) and an attention mechanism [5] (Attention), which all achieved excellent performance on semantic image seg-

| Method | Overall accuracy | Mean accuracy | Mean IoU |
|-----------------|------------------|---------------|--------------|
| SegNet [3] | 69.04 | 24.00 | 18.17 |
| FCN-8s [21] | 76.06 | 36.75 | 28.29 |
| DeepLabV2 [4] | 82.66 | 51.64 | 41.64 |
| Attention [5] | 83.43 | 54.39 | 42.92 |
| DeepLabV2 + SSL | 83.16 | 52.55 | 42.44 |
| Attention + SSL | 84.36 | 54.94 | 44.73 |

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

| Method | Overall accuracy | Mean accuracy | Mean IoU |
|-----------------|------------------|---------------|--------------|
| SegNet [3] | 69.10 | 24.26 | 18.37 |
| FCN-8s [21] | 76.28 | 37.18 | 28.69 |
| DeepLabV2 [4] | 82.89 | 51.53 | 41.56 |
| Attention [5] | 83.56 | 54.28 | 42.97 |
| DeepLabV2 + SSL | 83.37 | 52.53 | 42.46 |
| Attention + SSL | 84.53 | 54.81 | 44.59 |

Table 3: Comparison of human parsing performance with four state-of-the-art methods on the LIP test set.

mentations in different ways and have completely available codes. For a fair comparison, we train each method on our LIP training set for 30 epochs and evaluate on the validation set and the test set. For DeepLabV2, we use the VGG-16 model without dense CRFs. Following [5, 28], we use the standard intersection over union (IoU) criterion and pixel-wise accuracy for evaluation.

3.1. Overall performance evaluation

We begin our analysis by reporting the overall human parsing performance of each approach and summarize the results in Table. 2 and Table. 3. On the LIP validation set, among the four approaches, Attention [5] achieves the best result of 54.39% mean accuracy, benefited from the attention model that softly weights the multi-scale features. For mean IoU, Attention [5] performs best with 42.92%, while both FCN-8s [21] (28.29%) and SegNet [3] (18.17%) perform significantly worse. Similar performance is observed on the LIP test set. The interesting outcome of this comparison is that the achieved performance is substantially lower than the current best results on other segmentation benchmark such as PASCAL VOC [10]. This suggests that detailed human parsing due to the small parts and diverse fine-grained labels, is more challenging than object-level segmentation, which deserves more attention in the future.

3.2. Performance evaluation under different challenges

We further analyse the performance of each approach with respect to the following five challenging factors: occlusion, full-body, upper-body, head-missed and back-view (see Fig. 5). We evaluate the above four approaches on the LIP validation set which contains 4,277 images with occlusions, 5,452 full-body images, 793 upper-body images, 112 head-missed images and 661 back-view images. As expected, the performance varies when affected by different factors. Back-view is clearly the most challenging case. For example, the IoU of Attention [5] drops from 42.92%

| Method | hat | hair | gloves | sunglasses | u-clothes | dress | coat | socks | pants | jumpsuits | scarf | skirt | face | l-arm | r-arm | l-leg | r-leg | l-shoe | r-shoe | Bkg | Avg |
|-----------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|
| SegNet [3] | 26.60 | 44.01 | 0.01 | 0.00 | 34.46 | 0.00 | 15.97 | 3.59 | 33.56 | 0.01 | 0.00 | 0.00 | 52.38 | 15.30 | 24.23 | 13.82 | 13.17 | 9.26 | 6.47 | 70.62 | 18.17 |
| FCN-8s [21] | 39.79 | 58.96 | 5.32 | 3.08 | 49.08 | 12.36 | 26.82 | 15.66 | 49.41 | 6.48 | 0.00 | 2.16 | 62.65 | 29.78 | 36.63 | 28.12 | 26.05 | 17.76 | 17.70 | 78.02 | 28.29 |
| DeepLabV2 [4] | 57.94 | 66.11 | 28.50 | 18.40 | 60.94 | 23.17 | 47.03 | 34.51 | 64.00 | 22.38 | 14.29 | 18.74 | 69.70 | 49.44 | 51.66 | 37.49 | 34.60 | 28.22 | 22.41 | 83.25 | 41.64 |
| Attention [5] | 58.87 | 66.78 | 23.32 | 19.48 | 63.20 | 29.63 | 49.70 | 35.23 | 66.04 | 24.73 | 12.84 | 20.41 | 70.58 | 50.17 | 54.03 | 38.35 | 37.70 | 26.20 | 27.09 | 84.00 | 42.92 |
| DeepLabV2 + SSL | 58.41 | 66.22 | 28.76 | 20.05 | 62.26 | 21.18 | 48.17 | 36.12 | 65.16 | 22.94 | 14.84 | 19.37 | 70.01 | 50.45 | 53.39 | 37.59 | 36.96 | 26.29 | 26.87 | 83.67 | 42.44 |
| Attention + SSL | 59.75 | 67.25 | 28.95 | 21.57 | 65.30 | 29.49 | 51.92 | 38.52 | 68.02 | 24.48 | 14.92 | 24.32 | 71.01 | 52.64 | 55.79 | 40.23 | 38.80 | 28.08 | 29.03 | 84.56 | 44.73 |

Table 4: Performance comparison in terms of per-class IoU with four state-of-the-art methods on LIP validation set.

to 33.50%. The second most influential factor is the appearance of head. The scores of all approaches are much lower on head-missed images than the average score on the whole set. The performance also suffers a lot from occlusion. And the results of full-body images are the closest to the average level. By contrast, upper-body is relatively the easiest case, where fewer semantic parts are present and the part regions are usually larger. From the results, we can draw a conclusion that head(or face) is an important cue for the existing human parsing approaches. The probability of ambiguous results will increase if the head part disappears in the images or in the back-view. Moreover, the parts or clothes on the lower-body are more difficult than the ones on the upper-body because of the existence of small labels such as shoes or socks. In this case, body joint structure can play an effective role in guiding human parsing.

3.3. Per-class performance evaluation

In order to discuss and analyse each of the 20 labels in the LIP dataset in more detail, we further report the performance of per-class IoU on the LIP validation set, shown in Table. 4. We observe that the results with respect to labels with larger regions like face, upperclothes, coats and pants are much better than the ones on the small-region labels such as sunglasses, scarf and skirt. Attention [5] and DeepLabV2 [4] perform better on small labels thanks to the utilization of multi-scale features.

3.4. Visualization comparison

The qualitative comparisons of four approaches on our LIP validation set are visualized in Fig. 8. We display example parsing results of the five challenging factors scenarios. For the upper-body image(a) with slight occlusion, four approaches perform well with fewer errors. For the back-view image(b), all the four methods mistakenly label the right arm as left arm. The worst results appear when it comes to the head-missed image(c). SegNet [3] and FCN-8s [21] fail to recognize arms and legs, while DeepLabV2 [4] and Attention [5] have errors on the right and left of arms, legs and shoes. Furthermore, severe occlusion(d) also affects the performance a lot. Full-body is less challenging but the small objects in a full-body image(e) like shoes are also hard to be predicted precisely. Moreover, observed from (c) and (d), some of the results are unreasonable from the perspective of human body configuration(e.g. two shoes in one foot), because the existing approaches lack the consideration of body structures. In summary, human parsing is

more difficult than the general object segmentation. Particularly, human body structures should be paid more attention to strengthen the ability to predict human parts and clothes with more reasonable configurations. As a result, we consider connecting human parsing results and body joint structure to find out a better approach for human parsing.

4. Self-supervised Structure-sensitive Learning

4.1. Overview

As previously mentioned, a major limitation of the existing human parsing approaches is the lack of consideration of human body configuration, which is mainly investigated in the human pose estimation problem. The human parsing and pose estimation aim to label each image with different granularities, that is, pixel-wise semantic labeling versus joint-wise structure prediction. The pixel-wise labeling can address more detailed information while joint-wise structure provides more high-level structure. However, the results of state-of-the-art pose estimation models [32, 7] still have many errors. The predicted joints do not have high enough quality to guide human parsing compared with the joints extracted from parsing annotations. Moreover, the joints in pose estimation are not aligned with parsing annotations. For example, the arms are labeled as arms for parsing annotations only if they are not covered by any clothes, while the pose annotations are independent with clothes. To address these issues, in this work, we investigate how to leverage informative high-level structure cues to guide pixel-wise prediction. We propose a novel self-supervised structure-sensitive learning for human parsing, which introduces a self-supervised structure-sensitive loss to evaluate the quality of predicted parsing results from a joint structure perspective, as illustrated in Fig. 6.

Specifically, in addition to using the traditional pixel-wise annotations as the supervision, we generate the approximated human joints directly from the parsing annotations which can also guide human parsing training. In order to explicitly enforce the produced parsing results semantically consistent with the human joint structures, we treat the joint structure loss as a weight of segmentation loss which becomes our structure-sensitive loss.

4.2. Self-supervised Structure-sensitive Loss

Generally for the human parsing task, no other extensive information is provided besides the pixel-wise annotations. It means instead of using augmentative information,

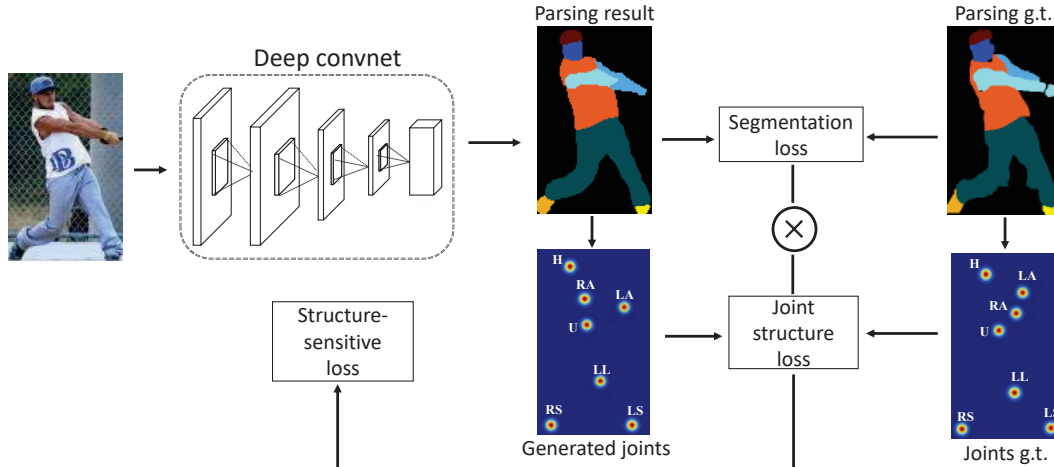


Figure 6: Illustration of our Self-supervised Structure-sensitive Learning for human parsing. An input image goes through parsing networks including several convolutional layers to generate the parsing results. The generated joints and joints ground truth represented as heatmaps are obtained by computing the center points of corresponding regions in parsing maps, including head (H), upper body (U), lower body (L), right arm (RA), left arm (LA), right leg (RL), left leg (LL), right shoe (RS), left shoe (LS). The structure-sensitive loss is generated by weighting segmentation loss with joint structure loss. For clear observation, here we combine nine heatmaps into one map.

we have to find a structure-sensitive supervision from the parsing annotations. As the human parsing results are semantic parts with pixel-level labels, we try to explore pose information contained in human parsing results. We define 9 joints to construct a pose structure, which are the centers of regions of head, upper body, lower body, left arm, right arm, left leg, right leg, left shoe and right shoe. The region of head are generated by merging parsing labels of hat, hair, sunglasses and face. Similarly, upper-clothes, coat and scarf are merged to be upper body, pants and skirt for lower body. The rest regions can also be obtained by corresponding labels. Some examples of generated human joints for different humans are shown in Fig. 7. Following [24], for each parsing result and corresponding ground truth, we compute the center points of regions to obtain joints represented as heatmaps for training more smoothly. Then we use Euclidean metric to evaluate the quality of the generated joint structures, which also reflect the structure consistency between the predicted parsing results and the ground truth. Finally, the pixel-wise segmentation loss is weighted by the joint structure loss, which becomes our structure-sensitive loss. Consequently the overall human parsing networks become self-supervised with the structure-sensitive loss.

Formally, given an image I , we define a list of joints configurations $C_I^P = \{c_i^p | i \in [1, N]\}$, where c_i^p is the heatmap of i -th joint computed according to the parsing result map. Similarly, $C_I^{GT} = \{c_i^{gt} | i \in [1, N]\}$, which is obtained from corresponding parsing ground truth. Here N is a variate decided by the human bodies in the input images which equals to 9 for a full-body image. For the joints missed in the image, we simply replace the heatmaps with maps filled with zeros. The joint structure loss is the Euclidean (L2) loss,

calculated as:

$$L_{\text{Joint}} = \frac{1}{2N} \sum_{i=1}^N \|c_i^p - c_i^{gt}\|_2^2 \quad (1)$$

The final structure-sensitive loss, denoted as $L_{\text{structure}}$, is the combination of the joint structure loss and the parsing segmentation loss, calculated as:

$$L_{\text{Structure}} = L_{\text{Joint}} \cdot L_{\text{Parsing}} \quad (2)$$

where L_{Parsing} is the pixel-wise softmax loss calculated based on the parsing annotations.

We phrase our learning framework “self-supervised” as this above structure-sensitive loss can be generated from existing parsing results without any extra information. Our self-supervised learning framework thus has excellent adaptability and extensibility which can be injected into any advanced networks to help incorporate rich high-level knowledge about human joints from a global perspective.

5. Experiments

5.1. Experimental Settings

Dataset: We evaluate the performance of our self-supervised structure-sensitive learning method on human parsing task on two challenging datasets. One is the public PASCAL-Person-part dataset with 1,716 images for training and 1,817 for testing, which pays attention to the human part segmentation annotated by [6]. Following [5, 28], the annotations are merge to be six person part classes and one background class which are Head, Torso, Upper / Lower arms and Upper / Lower legs. The other is our large-scale LIP dataset which is highly challenging with severe pose

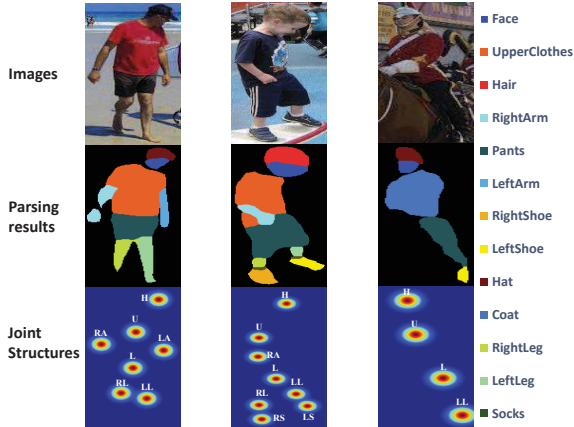


Figure 7: Some examples of self-supervised human joints generated from our parsing results for different bodies.

| Method | head | torso | u-arms | l-arms | u-legs | l-legs | Bkg | Avg |
|----------------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|
| DeepLab-LargeFOV [4] | 78.09 | 54.02 | 37.29 | 36.85 | 33.73 | 29.61 | 92.85 | 51.78 |
| HAZN [28] | 80.79 | 59.11 | 43.05 | 42.76 | 38.99 | 34.46 | 93.59 | 56.11 |
| Attention [5] | 81.47 | 59.06 | 44.15 | 42.50 | 38.28 | 35.62 | 93.65 | 56.39 |
| LG-LSTM [15] | 82.72 | 60.99 | 45.40 | 47.76 | 42.33 | 37.96 | 88.63 | 57.97 |
| Attention + SSL | 83.26 | 62.40 | 47.80 | 45.58 | 42.32 | 39.48 | 94.68 | 59.36 |

Table 5: Comparison of person part segmentation performance with four state-of-the-art methods on the PASCAL-Person-Part dataset [6].

complexity, heavily occlusions and body truncation, as introduced and analyzed in Section 3.

Network architecture: We utilize the publicly available model, Attention [5], as the basic architecture due to its leading accuracy and competitive efficiency. We also train a network based on DeepLabV2 [4], which employs re-purposed VGG-16 by atrous convolution, multi-scale inputs and atrous spatial pyramid pooling.

Training: We use the pre-trained models and networks settings provided by DeepLabV2 [4]. The scale of the input images is fixed as 321×321 for training networks based on Attention [5]. Two training steps are employed to train the networks. First, we train the basic network on our LIP dataset for 30 epochs, which takes about two days. Then we perform “self-supervised” strategy to fine-tune our model with structure-sensitive loss. We fine-tune the networks for roughly 20 epochs and it takes about one and a half days. We train all the models using stochastic gradient descent with a batch size of 10 images, momentum of 0.9, and weight decay of 0.0005. In the testing stage, one images takes 0.5 second on average.

Reproducibility: The proposed method is implemented by extending the Caffe framework. All networks are trained on a single NVIDIA GeForce GTX TITAN X GPU with 12GB memory. The code and models are available at https://github.com/Engineering-Course/LIP_SSL.

| Method | ATR | LIP | small | medium | large | 153 | 321 | 513 |
|-----------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|
| SegNet [3] | 15.79 | 21.79 | 16.53 | 18.58 | 18.18 | 16.92 | 18.37 | 16.44 |
| FCN-8s [21] | 34.44 | 32.28 | 22.37 | 29.41 | 28.09 | 14.52 | 15.55 | 16.25 |
| DeepLabV2 [4] | 48.64 | 43.97 | 28.77 | 40.74 | 43.02 | 36.49 | 37.59 | 37.28 |
| Attention [5] | 49.35 | 45.38 | 31.71 | 41.61 | 44.90 | - | - | - |
| DeepLab + SSL | 49.92 | 44.81 | 30.05 | 41.50 | 44.10 | 38.27 | 38.97 | 39.84 |
| Attention + SSL | 52.69 | 46.85 | 33.48 | 43.12 | 46.73 | - | - | - |

Table 6: Performance comparison in terms of mean IoU. Left: different test sets. Middle: different sizes of objects. Right: different single input sizes.

5.2. Results and Comparisons

We compare the proposed method with the strong baselines on the two public dataset.

PASCAL-Person-Part dataset [6]. Table. 5 shows the performance of our models and comparisons with four state-of-the-art methods on the standard intersection over union (IoU) criterion. Our method can significantly outperform four baselines, particularly. For example, our best model achieves 59.36%, 7.58% better than DeepLab-LargeFOV [4] and 2.97% better than Attention [5]. This large improvement demonstrates that our self-supervised strategy is significantly helpful for human parsing task.

LIP dataset: We report the results and the comparisons with four state-of-the-art methods on LIP validation set and test set in Table. 2 and Table. 3. On validation set, the proposed architecture can give a huge boost in average IoU: 3.09% better than DeepLabV2 [4] and 1.81% better than Attention [5]. On test set, our method also outperforms other baselines. This superior performance achieved by our method demonstrates the effectiveness of our self-supervised structure-sensitive learning, which incorporates the body joint structure into the pixel-wise prediction.

In Fig. 5, we show the results with respect to the different challenging factors on our LIP validation set. With our structure-sensitive loss, the performance of all kinds of types are improved, which demonstrates that human joint structure is helpful for human parsing task and our self-supervised learning is reasonable and efficient.

We further report per-class IoU on LIP validation set to verify the detailed effectiveness of our structure-sensitive loss, presented in Table. 4. With structure-sensitive loss, we achieved the best performance on almost all the classes. As observed from the reported results, structure-sensitive loss significantly improves the performance of the labels like arms, legs, and shoes, which demonstrates its ability to refine the ambiguous of left and right. Furthermore, the labels covering small regions such as sunglasses, socks, gloves, are predicted better with higher IoU. This improvement also demonstrates the effectiveness of structure-sensitive loss especially for small labels.

5.3. Qualitative Comparison

The qualitative comparisons of parsing results on the LIP validation set are visualized in Fig. 8. As can be observed

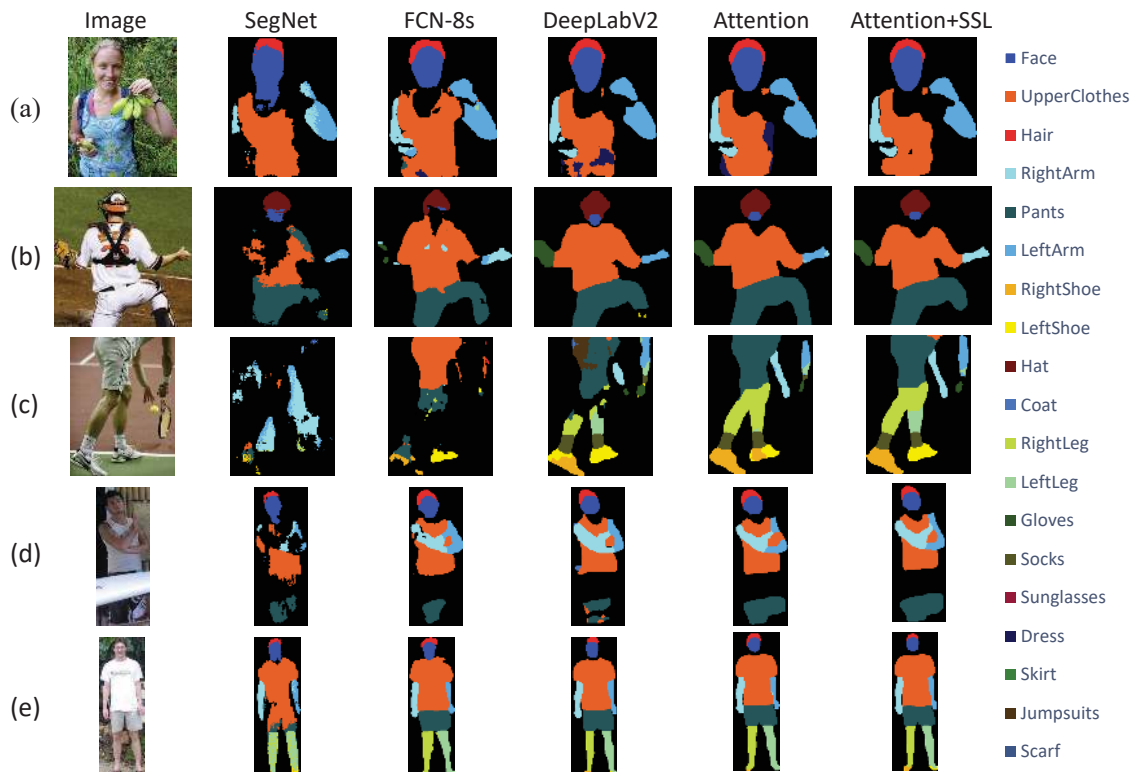


Figure 8: Visualized comparison of human parsing results on the LIP validation set. (a): The upper-body images. (b): The back-view images. (c): The head-missed images. (d): The images with occlusion. (e): The full-body images.

from these visualized comparisons, our self-learning structure outputs more semantically meaningful and precise predictions than other four methods despite the existence of large appearance and position variations. For example, observed from the full-body image(e), the small regions (e.g. left or right shoe) can be successfully segmented out by our method with structure-sensitive loss. Taking (b) and (c) for example, our approach can also successfully handle the confusing labels such as left arm versus right arm and left leg versus right leg. These regions with similar appearances can be recognized and separated by the guidance from joint structure information. For the most difficult head-missed image(c), the left shoe, right shoe and part of the left leg are excellently corrected by our approach. In general, by effectively exploiting self-supervised structure-sensitive loss, our approach outputs more reasonable results for confusing labels on the human parsing task.

5.4. Further experiments and analyses

For a better understanding of our LIP dataset, we evaluate the models trained on LIP and test on ATR [17] on the common categories, as reported in Table 6 (Left). In general, the performance on ATR are better than those on LIP because LIP dataset contains the instances with more diverse poses, appearance patterns, occlusions and resolution issues, which is more consistent with real-world situations.

Following MSCOCO dataset [18], we have done an em-

pirical analysis on different object sizes, i.e., small ($area < 153^2$), medium ($153^2 \leq area < 321^2$) and large ($area \geq 321^2$). The results of four baselines and the proposed SSL are reported in Table 6 (Middle). It can be observed that our SSL shows substantial superior performance for different sizes of objects. It further demonstrates the advantage of incorporating the structure-sensitive loss into parsing model.

To further research the influence of the input size, we perform an experiment over the scale of single input. The detailed analyses over different input sizes for all methods (except Attention [5] for its attention mechanism over scales) are presented in Table 6 (Right), which shows that our structure-sensitive learning is more robust for input size.

6. Conclusions

In this work, we presented “Look into Person (LIP)”, a large-scale human parsing dataset and a carefully designed benchmark to spark progress in human parsing. LIP contains 50,462 images, which are richly labeled with 19 semantic part labels. Taking advantage of our rich annotations, we performed detailed experimental analyses to identify the success and limitations of the leading human parsing approaches. Furthermore, we design a novel learning strategy, namely self-supervised structure-sensitive learning, to explicitly enforce the produced parsing results semantically consistent with the human joint structures.

References

- [1] M. Andriluka, L. Pishchulin, P. Gehler, and B. Schiele. 2d human pose estimation: New benchmark and state of the art analysis. In *CVPR*, June 2014.
- [2] P. Arbelaez, M. Maire, C. Fowlkes, and J. Malik. Contour detection and hierarchical image segmentation. *TPAMI*, 33(5):898–916, May 2011.
- [3] V. Badrinarayanan, A. Kendall, and R. Cipolla. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *arXiv preprint arXiv:1511.00561*, 2015.
- [4] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille. Semantic Image Segmentation with Deep Convolutional Nets and Fully Connected CRFs. In *ICLR*, 2015.
- [5] L.-C. Chen, Y. Yang, J. Wang, W. Xu, and A. L. Yuille. Attention to scale: Scale-aware semantic image segmentation. In *CVPR*, 2016.
- [6] X. Chen, R. Mottaghi, X. Liu, S. Fidler, R. Urtasun, et al. Detect what you can: Detecting and representing objects using holistic models and body parts. In *CVPR*, 2014.
- [7] X. Chen and A. Yuille. Articulated pose estimation by a graphical model with image dependent pairwise relations. In *NIPS*, 2014.
- [8] J. Dong, Q. Chen, X. Shen, J. Yang, and S. Yan. Towards unified human parsing and pose estimation. In *CVPR*, 2014.
- [9] J. Dong, Q. Chen, W. Xia, Z. Huang, and S. Yan. A deformable mixture parsing model with parselets. In *ICCV*, 2013.
- [10] M. Everingham, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman. The pascal visual object classes challenge 2010 (voc2010) results, 2010.
- [11] C. Gan, M. Lin, Y. Yang, G. de Melo, and A. G. Hauptmann. Concepts not alone: Exploring pairwise relationships for zero-shot video activity recognition. In *AAAI*, 2016.
- [12] C. Ionescu, D. Papava, V. Olaru, and C. Sminchisescu. Human3.6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. *TPAMI*, 36(7):1325–1339, jul 2014.
- [13] X. Liang, S. Liu, X. Shen, J. Yang, L. Liu, J. Dong, L. Lin, and S. Yan. Deep human parsing with active template regression. *TPAMI*, 2015.
- [14] X. Liang, S. Liu, Y. Wei, L. Liu, L. Lin, and S. Yan. Towards computational baby learning: A weakly-supervised approach for object detection. In *ICCV*, 2015.
- [15] X. Liang, X. Shen, D. Xiang, J. Feng, L. Lin, and S. Yan. Semantic object parsing with local-global long short-term memory. In *CVPR*, 2016.
- [16] X. Liang, Y. Wei, X. Shen, J. Yang, L. Lin, and S. Yan. Proposal-free network for instance-level object segmentation. *arXiv preprint arXiv:1509.02636*, 2015.
- [17] X. Liang, C. Xu, X. Shen, J. Yang, S. Liu, J. Tang, L. Lin, and S. Yan. Human parsing with contextualized convolutional neural network. In *ICCV*, 2015.
- [18] T. Lin, M. Maire, S. J. Belongie, L. D. Bourdev, R. B. Girshick, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft COCO: common objects in context. *CoRR*, abs/1405.0312, 2014.
- [19] S. Liu, X. Liang, L. Liu, X. Shen, J. Yang, C. Xu, L. Lin, X. Cao, and S. Yan. Matching-CNN Meets KNN: Quasi-Parametric Human Parsing. In *CVPR*, 2015.
- [20] Z. Liu, P. Luo, S. Qiu, X. Wang, and X. Tang. Deepfashion: Powering robust clothes recognition and retrieval with rich annotations. In *CVPR*, 2016.
- [21] J. Long, E. Shelhamer, and T. Darrell. Fully convolutional networks for semantic segmentation. In *CVPR*, 2015.
- [22] W. Lu, X. Lian, and A. Yuille. Parsing semantic parts of cars using graphical models and segment appearance consistency. In *BMVC*, 2014.
- [23] S. L. A. C. B. T. L. B. M. Hadi Kiapour, Xufeng Han. Where to buy it: matching street clothing photos in online shops. In *ICCV*, 2015.
- [24] T. Pfister, J. Charles, and A. Zisserman. Flowing convnets for human pose estimation in videos. In *ICCV*, 2015.
- [25] E. Simo-Serra, S. Fidler, F. Moreno-Noguer, and R. Urtasun. A High Performance CRF Model for Clothes Parsing. In *ACCV*, 2014.
- [26] J. Wang and A. Yuille. Semantic part segmentation using compositional model combining shape and appearance. In *CVPR*, 2015.
- [27] P. Wang, X. Shen, Z. Lin, S. Cohen, B. Price, and A. Yuille. Joint object and part segmentation using deep learned potentials. In *ICCV*, 2015.
- [28] F. Xia, P. Wang, L.-C. Chen, and A. L. Yuille. Zoom better to see clearer: Human part segmentation with auto zoom net. In *ECCV*, 2016.
- [29] F. Xia, J. Zhu, P. Wang, and A. Yuille. Pose-guided human parsing by an and/or graph using pose-context features. In *AAAI*, 2016.
- [30] K. Yamaguchi, M. Kiapour, and T. Berg. Paper doll parsing: Retrieving similar styles to parse clothing items. In *ICCV*, 2013.
- [31] K. Yamaguchi, M. Kiapour, L. Ortiz, and T. Berg. Parsing clothing in fashion photographs. In *CVPR*, 2012.
- [32] W. Yang, W. Ouyang, H. Li, and X. Wang. End-to-end learning of deformable mixture of parts and deep convolutional neural networks for human pose estimation. In *CVPR*, 2016.
- [33] H. Zhang, G. Kim, and E. P. Xing. Dynamic topic modeling for monitoring market competition from online text and image data. In *ACM SIGKDD*. ACM, 2015.
- [34] R. Zhao, W. Ouyang, and X. Wang. Unsupervised saliency learning for person re-identification. In *CVPR*, 2013.
- [35] S. Zheng, S. Jayasumana, B. Romera-Paredes, V. Vineet, Z. Su, D. Du, C. Huang, and P. Torr. Conditional random fields as recurrent neural networks. In *ICCV*, 2015.