Weakly and Semi Supervised Human Body Part Parsing via Pose-Guided Knowledge Transfer

Hao-Shu Fang[1], Guansong Lu[1], Xiaolin Fang[2]*, Jianwen Xie[3], Yu-Wing Tai[4], Cewu Lu[1]†
[1]Shanghai Jiao Tong University, China [2] Zhejiang University, China
[3]University of California, Los Angeles, USA [4] Tencent YouTu
fhaoshu@gmail.com sjtuluguansong@gmail.com fxlfang@gmail.com
jianwen@ucla.edu, yuwingtai@tencent.com lucewu@sjtu.edu.cn

Abstract

Human body part parsing, or human semantic part segmentation, is fundamental to many computer vision tasks. In conventional semantic segmentation methods, the ground truth segmentations are provided, and fully convolutional networks (FCN) are trained in an end-to-end scheme. Although these methods have demonstrated impressive results, their performance highly depends on the quantity and quality of training data. In this paper, we present a novel method to generate synthetic human part segmentation data using easily-obtained human keypoint annotations. Our key idea is to exploit the anatomical similarity among human to transfer the parsing results of a person to another person with similar pose. Using these estimated results as additional training data, our semi-supervised model outperforms its strong-supervised counterpart by **6** *mIOU on the PASCAL-Person-Part dataset [6], and we achieve state-of-the-art human parsing results. Our approach is general and can be readily extended to other object/animal parsing task assuming that their anatomical similarity can be annotated by keypoints. The proposed model and accompanying source code will be made* ***publicly available****.*

1. Introduction

The task of human body part parsing retrieves a semantic segmentation of body parts from the image of a person. Such pixel-level body part segmentations are not only crucial for activity understanding, but might also facilitate various vision tasks such as robotic manipulation [11], affordances reasoning [19] and recognizing human-object interactions [14]. In recent years, deep learning has been successfully applied to the human part parsing problem [4, 5, 32].

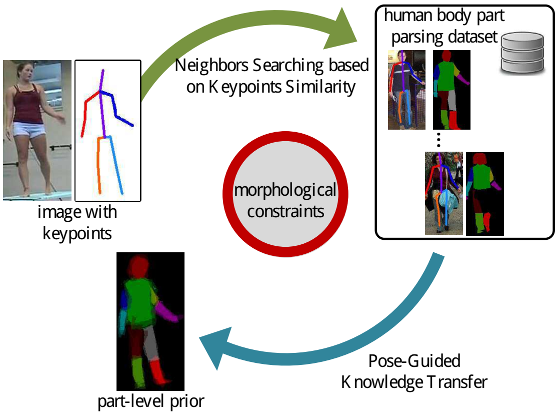
Neighbors Searching based on Keypoints Similarity
human body part parsing dataset
image with keypoints
morphological constraints
part-level prior
Pose-Guided Knowledge Transfer

Figure 1. Due to the morphological constraints, persons that share the same pose should have similar semantic part segmentations. For a person with only keypoints annotation, we search for persons in the human body part parsing dataset that have similar poses and then transfer their part parsing annotations to the target person. The transferred annotations form a strong part-level prior for the target person.

To fully exploit the power of deep convolutional neural networks, large-scale datasets are indispensable [9]. However, semantic labeling of body parts on a pixel-level is labor intensive. For the task of human body part parsing, the largest dataset [6] contains less than 2,000 labeled images for training, which is order of magnitudes less than the amount of training data in common benchmarks for image classification, semantic segmentation and keypoint estimation [9, 25, 12]. The small amount of data may lead to overfitting and degrade the performance in real world scenarios.

On the other hand, an abundance of human keypoints annotations [1] is readily available. Current state-of-the-art pose estimation algorithms [27] have also performed well in natural scene. The human keypoints encode structural information of the human body and we believe that such

*This work was done when Xiaolin Fang was an intern at MVIG lab of Shanghai Jiao Tong University.

†The corresponding author is Cewu Lu, email: lucewu@sjtu.edu.cn. Cewu Lu is also a member of SJTU-SenseTime lab and AI research institute of SJTU.

high-level human knowledge can be transfered to the task of human body part parsing.

However, despite the availability of keypoints annotations, few methods have investigated to utilize keypoints as augmented training data for human body part parsing. The main problem is that human keypoints are a sparse representation, while the human body part parsing requires an enormous amount of training data with dense pixel-wise annotations. Consequently, a end-to-end method which only relies on keypoint annotations as labels may not achieve high performance.

In this paper, we propose a novel approach to augment training samples for human parsing. Due to physical anatomical constraints, humans that share the same pose will have a similar morphology. As shown in Fig 1, given a person, we can use his/her pose annotation to search for the corresponding parsing results with similar poses. These collected parsing results are then averaged and form a strong part-level prior. In this way, we can convert the sparse keypoint representation into a dense body part segmentation prior. With the strong part-level prior, we combine it with the input image and forward them through a refinement network to generate an accurate part segmentation result. The generated part segmentation can be used as extra data to train a human parsing network.

We conduct exhaustive experiments on the PASCAL-Part dataset [6]. Our semi supervised method achieves state-of-the-art performance using a simple VGG-16 [30] based network, surpassing the performance of ResNet-101 [16] based counterpart trained on limited part segmentation annotations. When utilizing a model based on the deeper ResNet-101, our proposed method outperforms the state-of-the-art results by **3** mAP.

2. Related work

This paper is closely related to the following areas: semantic part segmentation, joint pose and body part estimation, and weakly supervised learning.

Human body part parsing. In this subtask of semantic segmentation, fully convolutional network (FCN) [26] and its variants [4, 5, 32] have demonstrated promising results. In [4], Chen *et al.* proposed atrous convolution to capture object features at different scales and they further combined the convolutional neural network (CNN) with a Conditional Random Field (CRF) to improve the accuracy. In [5], the authors proposed an attention mechanism that softly combines the segmentation predictions at different scales according to the context. To tackle the problem of scale and location variance, Xia *et al.* [32] developed a model that adaptively zoom the input image into the proper scale to refine the parsing results.

Another approach to human parsing is the usage of recurrent networks with long short term memory (LSTM) units[17]. The LSTM network can innately incorporate local and global spatial dependencies into their feature learning. In [23], Liang *et al.* proposed the local-global LSTM network to incorporate spatial dependencies at different distances to improve the learning of features. In [22], the authors proposed a Graph LSTM network to fully utilize the local structures (e.g., boundaries) of images. Their network takes arbitrary-shaped superpixels as input and propagates information from one superpixel node to all its neighboring superpixel nodes. To further explore the multi-level correlations among image regions, Liang *et al.* [21] proposed a structure-evolving LSTM that can learn graph structure during the optimization of LSTM network. These LSTM networks achieved competitive performance on human body part parsing.

Utilizing Pose for Human Parsing. Recent works try to utilize the human pose information to provide high-level structure for human body part parsing. Promising methods include pose-guided human parsing [34], joint pose estimation and semantic part segmentation [10, 20, 33] and self-supervised structure-sensitive learning [15]. These methods focus on using pose information to regularize part segmentation results, and they lie in the area of strong supervision. Our method differs from theirs in that we aim to transfer the part segmentation annotations to unlabeled data based on pose similarity and generate extra training samples, which emphasizes semi-supervision.

Weak Supervision for Semantic Segmentation. In [29, 7, 28, 8, 24, 2], the idea is to utilize weakly supervised methods for semantic segmentation. In particular, Chen *et al.* [7] proposed to learn segmentation priors based on visual subcategories. Dai *et al.* [8] harnessed easily obtained bounding box annotations to locate the object and generated candidate segmentation masks using unsupervised region proposal methods. Since bounding box annotations cannot generalize well for background (e.g. water, sky, grasses), Li *et al.* [24] further explored training semantic segmentation models using the sparse scribbles. In [2], Bearman *et al.* trained a neural network using supervision from a single point of each object. Under some time budget, the proposed supervised method may yield improved result compared to other weak supervised counterparts.

3. Our Method

3.1. Problem Definition

Our goal is to utilize pose information to weakly supervise the training of a human parsing network. Consider a semantic part segmentation problem where we have a dataset $\mathcal{D}_s = \{I_i, S_i, K_i\}_{i=1}^N$ of N labeled training examples. Let $I_i \in \mathrm{R}^{h\times w\times 3}$ denote an input image, $S_i \in \mathrm{R}^{h\times w\times u}$ de-

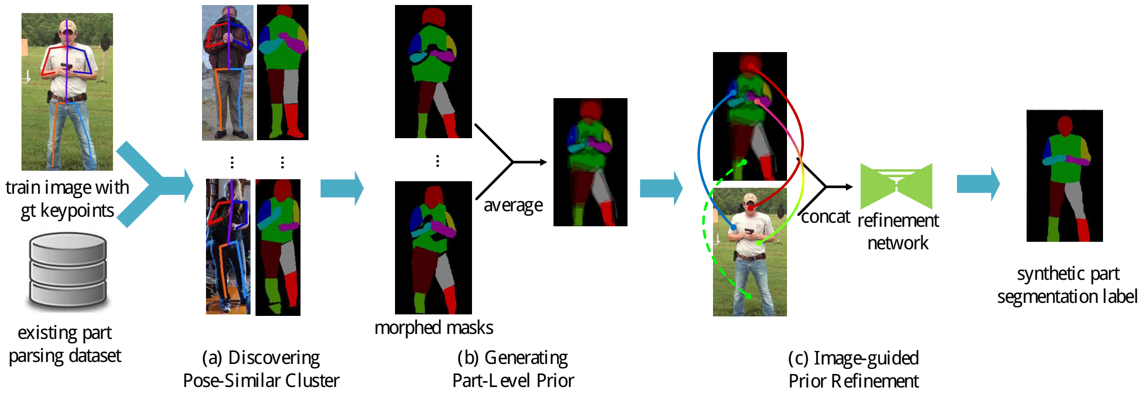
train image with gt keypoints
existing part parsing dataset
(a) Discovering Pose-Similar Cluster
morphed masks
average
(b) Generating Part-Level Prior
concat
refinement network
(c) Image-guided Prior Refinement
synthetic part segmentation label

Figure 2. Overview of our method. Given an image with keypoint annotations, we first search for the corresponding part segmentation with similar pose (a). Then, we apply a pose-guided morphing to the retrieved part segmentation masks and compute a part-level prior (b). A refinement network is applied to refine the prior based on local image evidence. The semantic part segmentation results can be used as extra training data to train a human parsing network without keypoint annotations (c). See text for more details.

notes its corresponding part segmentation label, $K_i \in \mathrm{R}^{v\times 2}$ denotes its keypoints annotation, and i is training example index. Each example contains at most u body parts and v keypoints. In practice, N is usually small since labeling the human semantic part is very labor intensive.

For a standard semantic part segmentation problem, the objective function can be written as:

$$\mathcal{E}(\Phi) = \sum_i \sum_j e[f_p^j(I_i;\Phi), S_i(j)], \tag{1}$$

where j is the pixel index of the image I_i, $S_i(j)$ is the semantic label at pixel j, $f_p(\cdot)$ is the fully convolutional network, $f_p^j(I_i;\Phi)$ is the per-pixel labeling produced by the network given parameters Φ, and $e[\cdot]$ is the per-pixel loss function.

Similarly, we can consider another dataset $\mathcal{D}_p = \{I_i, K_i\}_{i=1}^M$ of M examples with only keypoints annotations where $M \gg N$. Our goal is to generate part segmentations for all images in $\mathcal{D}_p$ and utilize these as additional training data when training a FCN by minimizing the objective function in Eqn. 1.

3.2. Overview

To utilize the keypoints annotations, we directly generate the pixel-wise part segmentations based on keypoints annotations. Given a target image I_t and keypoints annotation K_t where $(I_t, K_t) \in \mathcal{D}_p$, we first find a subset of images in $\mathcal{D}_s$ that have the most similar keypoints annotations (Sec. 3.3). Then, for the clustered images, we apply pose-guided morphing to their part segmentations to align them with the target pose (Sec. 3.4). The aligned body part segmentations are averaged to generate the part-level prior for the target image I_t. Finally, a refinement network is applied to estimate the body part segmentation of I_t (Sec. 3.5). The training method of our refinement network will be detailed in Sec. 3.6. The generated part segmentations can be used as extra training data for human part segmentation (Sec. 3.7). Figure 2 gives an overview of our method.

3.3. Discovering Pose-Similar Clusters

In order to measure similarity between different poses, we first need to normalize and align all the poses. We normalize the pose size by fixing their torsos to the same length. Then, the hip keypoints are used as the reference coordinate to align with the origin. We measure the Euclidean distances between K_t and every keypoint annotations in $\mathcal{D}_s$. The top-k persons in $\mathcal{D}_s$ with the smallest distances are chosen and form the pose-similar cluster, which serves as the basis for the following part-level prior generation step. The influence of k will be evaluated in Sec. 4.3.

Intuitively, given an image I_t with only keypoint annotations K_t, one may think of an alternative solution to obtain the part-parsing prior by solely morphing the part segmentation result of the person that has the closest pose to K_t. However, due to the differences between human bodies or possible occlusions, a part segmentation result with distinct boundary may not fit well to another one. Thus, instead of finding the person with the most similar pose, we find several persons that have similar poses and generate part-level prior by averaging their morphed part segmentation results. Such averaged part-level prior can be regarded as a probability map for each body part. It denotes the possibility for each pixel of whether it belongs to a body part based on real data distribution. In Sec. 4.3, we show that using the averaged prior achieves better performance than using only the parsing result of the closest neighbor.

3.4. Generating Part-Level Prior

The discovered pose-similar cluster forms a solid basis for generating the part-level prior. However, for each cluster, there are some inevitable intra-cluster pose variations, which makes the part parsing results misaligned. Thus, we

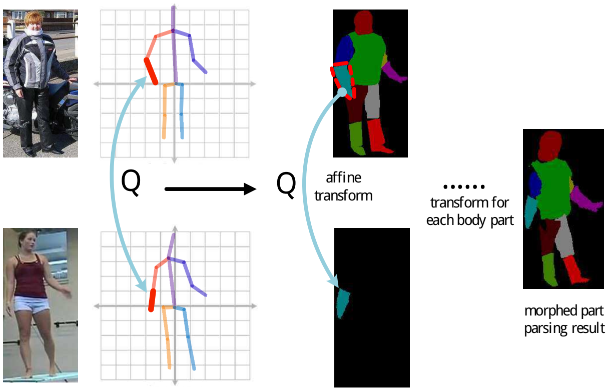

Figure 3. Pose-guided morphing for body part parsing. For each body part, we compute an affine transformation matrix $\boldsymbol{Q}$ according to the corresponding pose segment. The local body part is then transformed according to the estimated transformation matrix.

introduce the pose-guided morphing method.

For n persons in the same pose-similar cluster, let us denote their part parsing results as $\mathbb{S} = \{S_1, ..., S_n\}$, their keypoints annotations as $\mathbb{K} = \{K_1, ..., K_n\}$ and the morphed part parsing results as $\widetilde{\mathbb{S}} = \{\widetilde{S}_1, ..., \widetilde{S}_n\}$. By comparing the poses in $\mathbb{K}$ with the target pose K_t, we can compute the transformation parameters θ and then use them to transform $\mathbb{S}$ to obtain the morphed part parsing results $\widetilde{\mathbb{S}}$. We use the affine transformation to morph the part segmentations. This procedure can be expressed as:

$$\widetilde{\mathbb{S}} = \{T(S_i; \theta_i) \mid 1 \leq i \leq n \ , \ S_i \in \mathbb{S}\}, \tag{2}$$

where

$$\theta_i = g(K_t, K_i), K_i \in \mathbb{K},$$

$T(\cdot)$ is the affine transformation with parameters θ, and $g(\cdot)$ computes θ according to pose K_i and the target pose K_t.

For the part parsing annotations, we represent them as the combination of several binary masks. Each mask represent the appearance of a corresponding body part. The morphing procedure is conducted on each body part independently. Consider the left upper arm as an example. For the left upper arm segment $G_1 = \vec{\boldsymbol{x}}_1$ of pose K_1 and the same segment $G_2 = \vec{\boldsymbol{x}}_2$ of pose K_t, we have transformation relationship

$$\begin{pmatrix} \vec{\boldsymbol{x}}_1 \\ 1 \end{pmatrix} = \mathrm{Q} \begin{pmatrix} \vec{\boldsymbol{x}}_2 \\ 1 \end{pmatrix} = \begin{bmatrix} \boldsymbol{A} & \vec{\boldsymbol{b}} \\ \boldsymbol{0} & 1 \end{bmatrix} \begin{pmatrix} \vec{\boldsymbol{x}}_2 \\ 1 \end{pmatrix}, \tag{3}$$

where Q is the affine transformation matrix we need to calculate. Since both $\vec{\boldsymbol{x}}_1$ and $\vec{\boldsymbol{x}}_2$ are known, we can easily compute the result of $\boldsymbol{A}$ and $\vec{\boldsymbol{b}}$. Then, the morphed part segmentation mask can be obtained by

$$\begin{pmatrix} x_i^t \\ y_i^t \end{pmatrix} = \boldsymbol{A} \begin{pmatrix} x_i^s \\ y_i^s \end{pmatrix} + \vec{\boldsymbol{b}}, \tag{4}$$

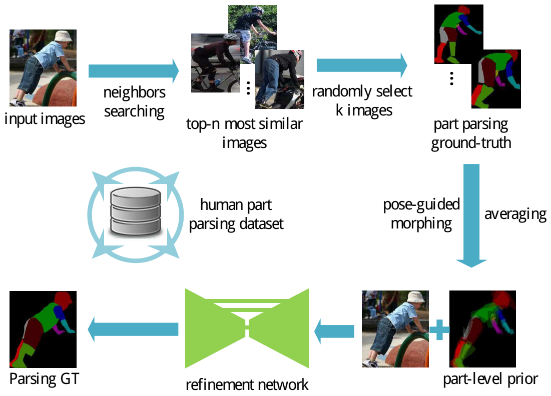

Figure 4. Training pipeline for our refinement network. We elaborate on the details in the text.

where $\{x_i^t, y_i^t\}$ and $\{x_i^s, y_i^s\}$ are the coordinates before and after transformation. Figure 3 illustrates our pose-guided morphing method.

After the pose-guided morphing, the transformed part segmentations $\widetilde{\mathbb{S}}$ are averaged to form the part-level prior

$$P_t = \frac{1}{n} \sum_{i=1}^{n} \widetilde{S}_i.$$

3.5. Prior Refinement

Finally, we feed forward our part-level prior through a refinement network together with the original image.

With a coarse prior, the search space for our refinement network is significantly reduced, and thus it can achieve superior results than directly making predictions from a single image input. The discriminative power of a CNN can eliminate the uncertainty at the body part boundary of the part parsing prior based on local image evidence, thus leading to high-quality parsing results. For each image I_t with body part prior P_t, we estimate the part segmentation result $\widehat{S}_t$ by

$$\widehat{S}_t = f_r(I_t, P_t; \Psi), \tag{5}$$

where f_r is the refinement network with parameters Ψ. In the next section, we will elaborate on the learning of Ψ. This estimated part segmentation result can be used as extra training data to train the FCN. The semi-supervised regime will be further discussed in Section 3.7.

3.6. Training of Refinement Network

Previous sections are conducted under the assumption that we have a well-trained refinement network. In this section, we will explain the training algorithm. Fig. 4 depicts a schematic overview of the proposed training pipeline.

The refinement network is a variant of "U-Net" proposed in [18], which is in the form of an auto-encoder network with skip connections. For such a network, the input will be progressively down-sampled until a bottleneck layer and then gradually up-sampled to restore the input size. To preserve the local image evidence, skip connections are introduced between each layer i and layer $n-i$, assuming the network has n layers in total. Each skip connection concatenates the feature maps at layer i with those in layer $n-i$. We refer readers to [18] for the detailed network structures. The input for our network is an image as well as a set of masks. Each mask is a heatmap ranging from 0 to 1 that represents the probability map for a specific body part. The output for this network is also a set of masks that have the same representation. The label is a set of binary masks indicating the appearance of each body part. Our objective function is the L1 distance between the output and the label. Both input and output size are set as 256×256.

To train the refinement network, we utilize the data in $\mathcal{D}_s$ with both semantic part segmentation annotations and pose annotations. Given an image $I_m \in \mathcal{D}_s$, similar to the pipeline for generating part-level prior in Sec. 3.4, we first generate the part parsing prior P_m given I_m. The only difference is that when discovering pose-similar cluster, we find n nearest neighbours each time and randomly pick k of them to generate part-level prior. This can be regarded as a kind of data augmentation to improve the generalization ability of the refinement network. The impact of n will be discussed in Sec. 4.3. Formally, with the part-level prior P_m and semantic part segmentation ground truth S_m for image I_m, we train our refinement network by minimizing the cost function:

$$\mathcal{E}(\Psi) = \sum_j \|S_m(j) - f_r^j(I_m, P_m; \Psi)\|_1. \tag{6}$$

3.7. Semi-Supervised Training for Parsing Network

In previous sections, we have presented our method to generate pixel-wise part segmentation labels based on keypoints annotations. Now we can train the parsing network for part segmentation in a semi-supervised manner. For our parsing network, we use the VGG-16 based model proposed in [5] due to its effective performance and simple structure. In this network, multi-scale inputs are applied to a shared VGG-16 based DeepLab model [4] for predictions. A soft attention mechanism is employed to weight the outputs of the FCN over scales. The training for this network follows the standard process of optimizing the per-pixel regression problem, which is formulated as Eqn. 1. For the loss function, we use the multinomial logistic loss. During training, the input image is resized and padded to 320×320.

We consider updating the parameter Φ of network f_p by minimizing the objective function in Eqn. 1 on $\mathcal{D}_s$ with ground truth labels and $\mathcal{D}_p$ with generated part segmentation labels.

4. Experiments

In this section, we first introduce related datasets and implementation details in our experiments. Then, we report our results and comparisons with state-of-the-art performance. Finally, we conduct extensive experiments to validate the effectiveness of our proposed semi-supervision method.

4.1. Datasets

Pascal-Person-Part Dataset [6] is a dataset for human semantic part segmentation. It contains 1,716 images for training and 1,817 images for testing. The dataset contains detailed pixel-wise annotations for body parts, including hands, ears, *etc*. The keypoint annotations for this dataset have been made available by [33].

MPII [1] is a challenging benchmark for person pose estimation. It contains around 25K images and over 40K people with pose annotations. The training set consists of over 28K people and we select those with full body annotations as extra training data for human part segmentation. After filtering, there remain 10K images with single person pose annotations.

Horse-Cow Dataset [31] is a part segmentation benchmark for horse and cow images. It contains 294 training images and 227 testing images. The keypoint annotations for this dataset are provided by [3]. In addition, 317 extra keypoint annotations are provided by [3], which have no corresponding pixel-wise part parsing annotations. We take these keypoint annotations as extra training data for our network.

4.2. Implementation Details

In our experiments, we merge the part segmentation annotations in [6] to be Head, Torso, Left/Right Upper Arms, Left/Right Lower Arms, Left/Right Upper Legs and Left/Right Lower Legs, resulting in 10 body parts. The pose-guided morphing is conducted on the mask of each body part respectively. In order to be consistent with previous works, after the morphing, we merge the masks of each left/right pair by max-pooling and get six body part classes.

For our semi-supervision setting, we first train the refinement network and then fix it to generate synthetic part segmentation data using keypoints annotations in [1]. Note that for simplicity, we only synthesize part segmentation labels for the single person case, and it would easy to extend to a multi-person scenario. To train our refinement network with single person data, we crop those people with at least upper body keypoints annotations in the training list of the part parsing dataset [6] and get 2,004 images with a single person and corresponding part segmentation label. We randomly sample 100 persons as validation set for the refinement network to set hyper-parameters.

To train the refinement network, we apply random jittering by first resizing input images to 286×286 and then randomly cropping to 256×256. The batch size is set to

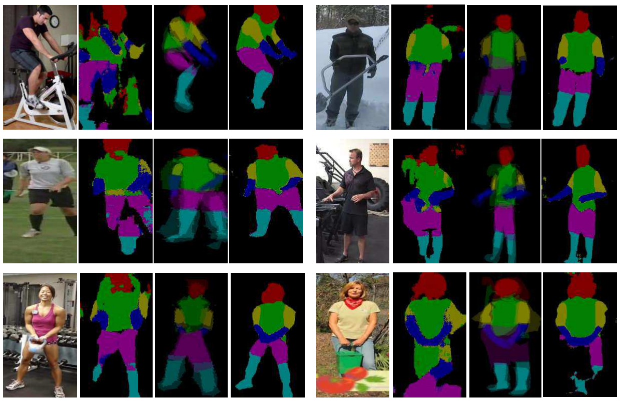

Figure 5. Qualitative comparison of the refinement network trained with and without part-level prior. For each image group, from left to right: input image, FCN prediction without prior, corresponding part-level prior for the input image, prediction from refinement network trained with prior.

1. We use the Adam optimizer with an initial learning rate of 2e$-$4 and β_1 of 0.5. To train our parsing network, we follow the same setting as [5] by setting the learning rate as 1e-3 and decayed by 0.1 after 2,000 iterations. The learning rate for the last layer is 10 times larger than previous layers. The batch size is set to 30. We use the SGD solver with a momentum of 0.9 and weight decay of 5e$-$4. All experiments are conducted on a single Nvidia Titan X GPU. Our refinement network is trained from scratch for 200 epochs, which takes around 18 hours. The parsing network is initialized with a model pre-trained on COCO dataset [25] which is provided by the author. We train it for 15K iterations, and it takes around 30 hours for training.

4.3. Results and Comparisons

Evaluation of Annotation Number. We first compare the performance of the parsing network trained with a different number of annotations on validation set and the results are reported in Table 1. When all annotations of semantic part segmentations are used, the result is the original implementation of the baseline model [5], and our reproduction is slightly (0.5 mIOU) higher. Then we gradually add the data with keypoint annotations. As can be seen, the performance increases in line with the number of keypoint annotations. This suggests that our semi supervision method is effective and scalable. Fig. 6 gives some qualitative comparisons between the predictions of the baseline model and our semi-supervised model.

supervision	mask anno.#	keypoints anno.#	mIoU
full	7K	-	56.89
semi	7K	4K	59.60
	7K	7K	61.44
	7K	10k	62.60

Table 1. Results on **PASCAL-Person-Part** dataset. In the "supervision" column, "full" means all training samples are with segmentation mask annotations, "semi" means mixtures of mask annotations and keypoints annotations.

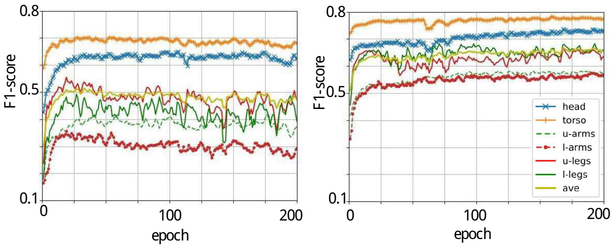

Figure 7. Validation accuracy of refinement network with and without part-level prior. The left curve is the case trained without prior and the right one is the case with prior. The performance of refinement network is much better with the part-level prior.

strategy	cluster size *k*	pool size *n*	mIoU
skeleton label map [33]	-	-	58.26
part-level prior, w/o aug	1	-	60.10
	3	-	61.78*
	5	-	61.26
	7	-	60.32
part-level prior, w aug	3	5	62.60

Table 2. Performances of different prior generation strategies on **PASCAL-Person-Part** dataset.

Significance of Part-level Prior. We evaluate the significance of the part-level prior. First, we compare the performance of refinement network trained with or without part-level prior on the cropped single person part segmentation validation set. Without prior, the refinement network is a regular FCN trained with limited training data. The validation accuracy for both cases is shown in Fig. 7. As we can see, the accuracy of the refinement network with part-level prior is much higher than the case without prior. In Fig. 5, we visualize some predictions from our refinement network trained with and without part-level priors, and the corresponding priors are also visualized.

Prior Generation Strategy. During our training process, the quality of part-level prior is important and directly affects the refined part segmentation results. We compare different prior generation strategies and report the final results in Table 2. We first explore using the skeleton label map [33]as prior, which draws a stick with width 7 between neighboring joints, and the result is 58.26 mIOU. Comparing to this method, our proposed part-level prior has a considerable improvement, which indicates the importance of knowledge transfer during prior generation.

For part-level prior, we compare the impact of the size k of pose-similar cluster. As aforementioned in Sec. 3.3, if we only choose the person with the nearest pose and take his/her morphed part parsing result as our prior, the performance is limited. But if we choose too many people to generate our part-level prior, the quality of the final re-

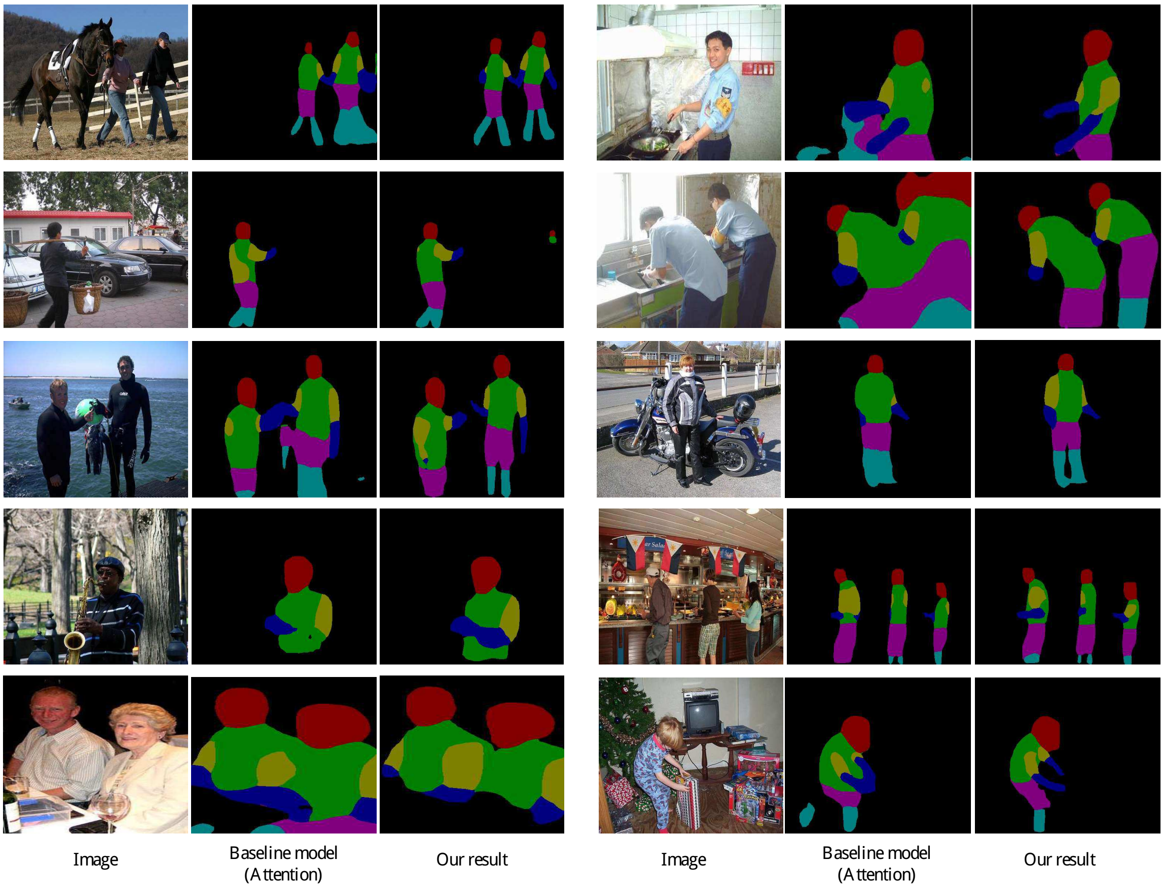

Figure 6. Qualitative comparison on the PASCAL-Person-Part dataset between the baseline model and our semi-supervised model.

sults would also decline. We claim that this is because our part segmentation dataset is small and the intra-cluster part appearance variance would increase as the cluster size increases. Then, we explore the influence of adding data augmentation during the training of refinement network. As we can see, randomly sample 3 candidates to generate part-level prior from a larger pool with size 5 is beneficial for training since it increases the sample variances. For the remaining experiments, we set k as 3 and n as 5.

Comparisons with State-of-the-Art. In Table 3, we report comparisons with the state-of-the-art results on the Pascal-Person-Part dataset [6]. With additional training data generated by our refinement network, our VGG16 based model achieves the state-of-the-art performance. Although we focus on exploiting human keypoints annotations as extra supervision, our method applies to other baseline models and can benefit from other parallel improvements. To prove that, we replace our baseline model with ResNet-101 [16] based DeepLabv2 [4] model and follow the same training setting. Without additional keypoints annotations, this baseline achieves 59.6 mAP. Under our semi-supervised training, this model achieves **64.28** mIOU. Note that this result is obtained by single scale testing. By performing multi-scale testing (scale = 0.5, 0.75, 1), we can further achieve **67.6** mIOU, which outperforms previous best result by **3** mIOU.

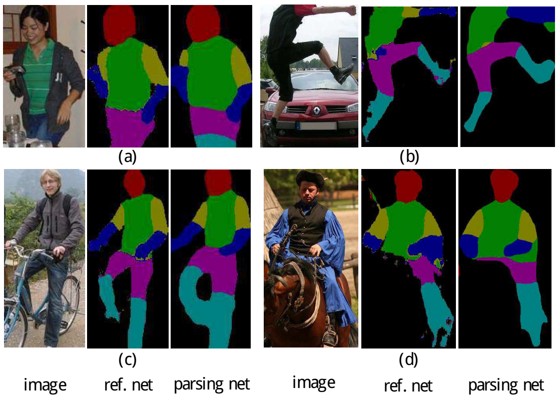

Figure 8. Qualitative comparison between the predictions of the refinement network and the parsing network. “ref.net” denotes the predictions of the refinement network, and “parsing net” denotes the predictions of the parsing network.

Comparisons between two Networks. To see how the refinement network can assist the training of the parsing

Method	head	torso	u-arms	l-arms	u-legs	l-legs	Bkg	Avg
DeepLab-LargeFOV-CRF [4]	80.13	55.56	36.43	38.72	35.50	30.82	93.52	52.95
Attention [5]	81.47	59.06	44.15	42.50	38.28	35.62	93.65	56.39
HAZN [32]	80.79	80.76	45.65	43.11	41.21	37.74	93.78	57.54
Graph LSTM [23]	82.69	62.68	46.88	47.71	45.66	40.93	94.59	60.16
Structure-evolving LSTM [21]	82.89	67.15	51.42	48.72	51.72	45.91	97.18	63.57
Joint (VGG-16, +ms) [33]	80.21	61.36	47.53	43.94	41.77	38.00	93.64	58.06
Joint (ResNet-101, +ms) [33]	85.50	67.87	54.72	54.30	48.25	44.76	95.32	64.39
Ours (VGG-16)	84.06	67.03	51.66	50.15	45.33	44.26	95.73	62.60
Ours (ResNet-101)	84.83	68.64	53.11	53.01	48.40	46.76	95.22	64.28
Ours (ResNet-101, +ms)	**87.15**	**72.28**	**57.07**	**56.21**	**52.43**	**50.36**	**97.72**	**67.60**

Table 3. Comparison of semantic object parsing performance with several state-of-the-art methods on the PASCAL-Person-Part dataset [6]. "+ms" denotes testing with multi-scale inputs. Note that we only perform single scale testing for our VGG-16 entry since the base network [5] has explicitly utilized multi-scale features.

network, we visualize some predictions of both networks in Fig. 8. In this experiment, the parsing network we use has already been trained under the semi supervision setting. As shown in Fig. 8, due to the guidance of the strong part-level prior, the refinement network makes fewer mistakes on the structure predictions (e.g., the upper legs in (a) and the upper arms in (b)), and produces tighter masks (e.g., the legs in (c)). These improvements obtained by using the part-level prior will be transferred to the parsing network during the semi-supervised training. On the other hand, by leveraging a large number of training data, the parsing network can figure out which are important or irrelevant features and produce predictions that are less noisy (e.g.,(b) and (d)).

Training on Pose Estimation Results. Since single person pose estimation is mature enough to be deployed, what if we replace the ground-truth keypoint annotations with pose estimation results? To answer that, we use the pose estimator [13] trained on MPII dataset to estimate human poses in COCO dataset [25]. Same as previous, we crop those people with full body annotations and collect 10K images. The semi-supervised result achieves **61.8** mIOU, which is on par with the results trained on ground-truth annotations. It shows that our system is robust to noise and suggests a promising strategy to substantially improve the performance of human semantic part segmentation without extra costs.

Extension to other Categories. To show the potential of extending our method to other categories, we also perform experiments on the Horse-Cow Dataset [31]. The results are reported in Table 4. Our baseline model, which is the attention model [5], has an mIOU of 71.55 for the Horse and an IOU of 68.84 for the Cow. By leveraging the keypoint annotations provided by [3], we gain improvements of 3.14 and 3.39 mIOU for Horse and Cow respectively. The consistent improvements across different categories indicate that our method is general and applicable to other segmentation tasks where their anatomical similarity can be annotated by keypoints. Finally, by replacing our baseline model with the deeper ResNet-101 based model, we achieve the most state-of-the-art result on the Horse-Cow dataset, yielding the final performances of 76.99 and 74.22 mIOU for these two categories respectively.

			Horse			
Method	Bkg	head	body	leg	tail	Avg
HAZN [32]	90.94	70.75	84.49	63.91	51.73	72.36
Graph LSTM [23]	91.73	72.89	86.34	69.04	53.76	74.75
Structure-evolving LSTM [21]	92.51	74.89	87.55	71.93	**57.45**	76.87
Attention [5]	90.48	68.91	83.34	64.20	50.74	71.55
Ours(VGG-16)	91.62	72.75	87.24	69.52	52.32	74.69
Ours(ResNet-101)	**93.39**	**75.63**	**88.39**	**72.61**	54.95	**76.99**

			Cow			
Method	Bkg	head	body	leg	tail	Avg
HAZN [32]	90.71	75.18	83.33	57.42	29.37	67.20
Graph LSTM [23]	91.54	73.88	85.92	63.67	35.22	70.05
Structure-evolving LSTM [21]	92.88	77.75	87.91	67.60	**42.86**	73.80
Attention [5]	91.06	74.33	84.38	60.60	33.85	68.84
Ours(VGG-16)	92.27	78.36	88.20	64.76	37.58	72.23
Ours(ResNet-101)	**93.70**	**80.30**	**89.53**	**66.92**	40.65	**74.22**

Table 4. Comparison of object parsing performance with three state-of-the-art methods over the Horse-Cow object parsing dataset [31]. We also list the performance of Attention [5], which is the baseline model for our VGG-16 entry.

5. Conclusion

In this paper, we propose a novel strategy to utilize the keypoint annotations to train deep network for human body part parsing. Our method exploits the constraints on biological morphology to transfer the part parsing annotations among different persons with similar poses. Experimental results show that it is effective and general. By utilizing a large number of keypoint annotations, we achieve the most state-of-the-art result on human part segmentation.

Acknowledgement: This work is supported in part by the National Natural Science Foundation of China under Grants 61772332. This work is also support in part by Sensetime Ltd

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*, 2014. 1, 5

[2] A. Bearman, O. Russakovsky, V. Ferrari, and L. Fei-Fei. What's the point: Semantic segmentation with point supervision. In *ECCV*, 2016. 2

[3] L. Bourdev, S. Maji, and J. Malik. Detection, attribute classification and action recognition of people using poselets (in submission). In *T'PAMI*. 5, 8

[4] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. 2015. 1, 2, 5, 7, 8

[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. 1, 2, 5, 6, 8

[6] X. Chen, R. Mottaghi, X. Liu, S. Fidler, R. Urtasun, and A. Yuille. Detect what you can: Detecting and representing objects using holistic models and body parts. In *CVPR*, 2014. 1, 2, 5, 7, 8

[7] X. Chen, A. Shrivastava, and A. Gupta. Enriching visual knowledge bases via object discovery and segmentation. In *CVPR*, 2014. 2

[8] J. Dai, K. He, and J. Sun. Boxsup: Exploiting bounding boxes to supervise convolutional networks for semantic segmentation. In *ICCV*, 2015. 2

[9] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. In *CVPR*, 2009. 1

[10] J. Dong, Q. Chen, X. Shen, J. Yang, and S. Yan. Towards unified human parsing and pose estimation. In *CVPR*, 2014. 2

[11] S. Ekvall and D. Kragic. Interactive grasp learning based on human demonstration. In *ICRA*, 2004. 1

[12] M. Everingham, S. M. A. Eslami, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The pascal visual object classes challenge: A retrospective. *IJCV*, 111:98–136, 2015. 1

[13] H.-S. Fang, S. Xie, Y.-W. Tai, and C. Lu. RMPE: Regional multi-person pose estimation. In *ICCV*, 2017. 8

[14] G. Gkioxari, R. Girshick, and J. Malik. Actions and attributes from wholes and parts. In *ICCV*, 2015. 1

[15] K. Gong, X. Liang, X. Shen, and L. Lin. Look into person: Self-supervised structure-sensitive learning and a new benchmark for human parsing. *arXiv preprint arXiv:1703.05446*, 2017. 2

[16] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 2, 7

[17] S. Hochreiter and J. Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997. 2

[18] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros. Image-to-image translation with conditional adversarial networks. *arxiv*, 2016. 4, 5

[19] H. S. Koppula, R. Gupta, and A. Saxena. Learning human activities and object affordances from rgb-d videos. *The International Journal of Robotics Research*, 32(8):951–970, 2013. 1

[20] L. Ladicky, P. H. Torr, and A. Zisserman. Human pose estimation using a joint pixel-wise and part-wise formulation. In *ICCV*, 2013. 2

[21] X. Liang, L. Lin, X. Shen, J. Feng, S. Yan, and E. P. Xing. Interpretable structure-evolving lstm. *arXiv preprint arXiv:1703.03055*, 2017. 2, 8

[22] X. Liang, X. Shen, J. Feng, L. Lin, and S. Yan. Semantic object parsing with graph lstm. In *ECCV*, 2016. 2

[23] 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. 2, 8

[24] D. Lin, J. Dai, J. Jia, K. He, and J. Sun. Scribblesup: Scribble-supervised convolutional networks for semantic segmentation. In *ICCV*, 2016. 2

[25] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft coco: Common objects in context. In *ECCV*, 2014. 1, 6, 8

[26] J. Long, E. Shelhamer, and T. Darrell. Fully convolutional networks for semantic segmentation. In *CVPR*, 2015. 2

[27] A. Newell, K. Yang, and J. Deng. Stacked hourglass networks for human pose estimation. In *ECCV*, 2016. 1

[28] G. Papandreou, L.-C. Chen, K. Murphy, and A. L. Yuille. Weakly-and semi-supervised learning of a dcnn for semantic image segmentation. *arXiv preprint arXiv:1502.02734*, 2015. 2

[29] D. Pathak, E. Shelhamer, J. Long, and T. Darrell. Fully convolutional multi-class multiple instance learning. *arXiv preprint arXiv:1412.7144*, 2014. 2

[30] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 2

[31] J. Wang and A. L. Yuille. Semantic part segmentation using compositional model combining shape and appearance. In *CVPR*, 2015. 5, 8

[32] F. Xia, P. Wang, L.-C. Chen, and A. L. Yuille. Zoom better to see clearer: Human and object parsing with hierarchical auto-zoom net. In *ECCV*, 2016. 1, 2, 8

[33] F. Xia, P. Wang, X. Chen, and A. Yuille. Joint multi-person pose estimation and semantic part segmentation. *arXiv preprint arXiv:1708.03383*, 2017. 2, 5, 6, 8

[34] F. Xia, J. Zhu, P. Wang, and A. L. Yuille. Pose-guided human parsing by an and/or graph using pose-context features. In *AAAI*, 2016. 2