POSE2BODY: POSE-GUIDED HUMAN PARTS SEGMENTATION

Zhong Li[1†], *Xin Chen*[2,3,4†], *Wangyiteng Zhou*[2,3,4], *Yingliang Zhang*[2,5], *Jingyi Yu*[2]

[1]University of Delaware, Newark, USA
[2]School of Information Science and Technology, ShanghaiTech University, China
[3]Shanghai Institute of Microsystem and Information Technology, Chinese Academy of Sciences, China
[4]University of Chinese Academy of Sciences, China [5]DGene Inc, Shanghai, China

ABSTRACT

Reliable human parts segmentation on 2D images plays an important role in many human-centric computer vision tasks. While significant achievements have been made on human pose estimation, the performance on human parts segmentation remains low. In this paper, we present a novel technique that we call Pose2Body that robustly conducts human parts segmentation based on the pose estimation results. We partition an image into superpixels and set out to assign a segment label to each superpixel most consistent with the pose. We design special feature vectors for every superpixel-label assignment as well as superpixel-superpixel pairs and model optimal labeling as to solve for a conditional random field (CRF). Comprehensive experiments show that our technique achieves substantial improvements over the state-of-the-art solutions.

Index Terms— Semantic labeling; human parts segmentation; pose estimation; conditional random field

1. INTRODUCTION

Semantic segmentation, also known as image labeling or scene parsing, aims to assign semantic labels to each pixel on the image. Successful solutions can benefit numerous applications ranging from scene understanding to 3D reconstruction. Most recently, the emphasis has shifted to the more specific task of human part parsing into anatomically meaningful components - head, neck, trunk, and upper and lower limbs. Such human-centric analysis enables more reliable person identification, video surveillance, virtual clothes fitting, etc.

Human parts semantic segmentation is also closely related to human pose estimation which has achieved significant improvements over the past few years with the help of deep convolutional neural networks. Latest solutions can simultaneously achieve high accuracy and reliability. For example, the RMPE method [1] employ a novel regional multi-person pose estimation framework to and can achieve 82.1% mAP accuracy on MPII dataset. In contrast, the accuracy in human parts segmentation still falls short: the state-of-the-art solutions [2, 3, 4, 5] achieve an accuracy range from 50 to 58 mIOU on VOC PASCAL-Human-Part dataset and are still sensitive to occlusion, person size and unusual pose.

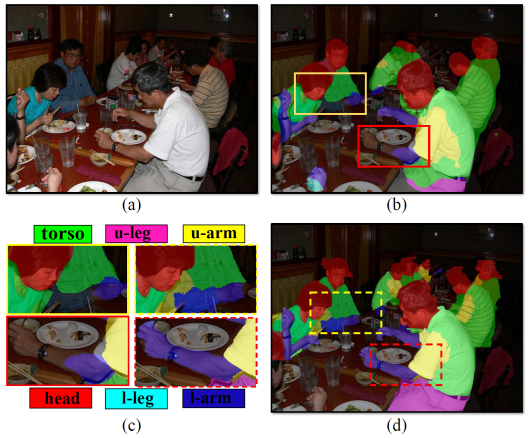

Fig. 1. Our semantic body parts segmentation compared with FCN [4]. (a) is input image. (b) is FCN result. (d) is our final result guided by pose estimation. (c) is the close-up view comparison.

We observe that human part segmentation and pose estimation are two complementary problems. On one hand, if we correctly conduct per-pixel semantic labeling, we can group the pixels of the same label to form the skeleton and hence pose. On the other, pose estimation provides not only valuable priors on the pixel labeling but also important connectivity, e.g., the arm should be connected to the forearm as well as to the trunk. We therefore present a joint pose and human part segmentation scheme that can conduct multi-person segmentation.

Given an estimation of human pose that consists of a set of vertices (joints) and skeleton edges (segments), we partition the image into superpixels and set out to assign a segment label to each superpixel. We design a 5D feature vector for each superpixel-label assignment to account for the spatial relationship between the superpixel location and the skeleton. Next, we compute the optimal labeling by considering

† These authors contributed to the work equally.

all pairs of superpixels via a conditional random field (CRF). Specifically, for each pair of superpixels and their corresponding labels, we compute a 11D feature vector (5D for each superpixel and the last dimension for computing the likelihood of the two superpixels sharing a joint). We train a group of classifiers on the feature vector using logistic regression and then measure the likelihood of pair-wise superpixel semantic assignments.

To formulate the final energy function, we use the latest RMPE [1] to obtain pose estimation and the deeplab semantic segmentation framework [2] to obtain per-pixel semantic labeling stored as a score map as initialization. The final energy function consists of two terms, the first unary term for measuring deviations of the proposed label to the semantic score map whereas the second pairwise term directly uses the outputs from the trained classifiers. Finally, we solve for this CRF using integer linear programming. We conduct extensive experiments on VOC PASCAL-Person-Part dataset. In particular, we show that on the above dataset, our technique achieves substantial improvements over the state-of-the-art. Fig. 1 shows our segmentation result compare with baseline method FCN [4].

2. RELATED WORKS

Compared with human pose estimation, human parts segmentation has adopted completely different solution routes. Traditional techniques have long used the graphical model to solve for human-clothes parsing where the goal is to label each image pixel as either a semantic apparel label or human body [6]. These techniques are sensitive to occlusions and movements. Xia et al. [7] guided human parts parsing using pose-guided segmentation proposals that can handle complex movements.

Same as pose estimation, recent successful approaches unanimously adopted the fully convolution neural networks (FCN) [4]. Techniques such as the DeepLab models [8] made use of the fully connected pairwise Conditional Random Field (CRF) as an auxiliary post-processing step to refine the initial estimation. Wang et al. [9] adopted a simpler and faster solution by leveraging foreground/background separation. However, their technique are sensitive to local ambiguity due to the FCN's inherent invariance property. For example, legs can be incorrectly labeled as arms and background can be labeled as legs.

Improved FCN-type approaches employ scale variations (different size of humans in the image). Chen et al. [5] used multiple scaled inputs to conduct results consolidation and sought to find the optimal scale configuration. [10] addressed the scale issue through "zoom and refine" to handle smaller body parts that are traditionally challenging in classical deep network solutions. Most recently, [2] applied the Deep Residual Net to improve the performance by effectively exploiting contextual information on individual pixels. However, by far the performance on body parts segmentation is still inferior to pose estimation.

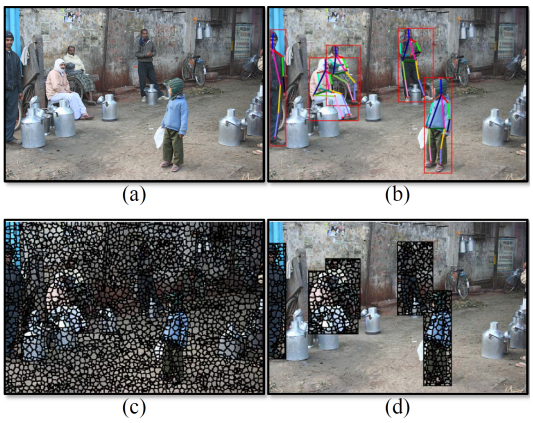

Fig. 2. Superpixel generation. (a) is input image. (b) shows several bounding boxes detected by human detector [11]. (c) is an initial superpixel result. We combine (b) and (c) to get the final superpixel patches in (d).

Our approach leverages both advances on pose estimation and human parts segmentation. Closest to our work is the approach by Xia et al [3] that refines semantic parts with joint estimation. Specifically, they directly input the joint label map and initial semantic part score map to a FCN-type neural network. We in contrast conduct a superpixel labeling and model the optimization problem as to solve for a CRF. In particular, we explicitly measure pair-wise label differences between superpixels based on a new class of feature descriptors and we show this new formulation significantly improves the performance.

3. HUMAN PARTS LABELING

Our approach assumes human poses on 2D images are known and sets out to label every pixel to a specific human part. For poses, we adopt 14 joint representations composed of forehead, neck, left/right shoulder, left/right elbow, left/right wrist, left/right waist, left/right knee and left/right ankle. We follow the same semantic human part labeling notation that partitions human body into $\mathbf{C} = 7$ labels: head, torso, upper arm, lower arm, upper leg, lower leg, and background (bg).

3.1. Problem Formulation

Before proceeding, we first clarify our notations. Given an image I, our goal is to output a pixel-wise labeling for each human part. Instead of conducting per-pixel estimation, we first partition the image into superpixels. Specifically, we adopt the Faster Mask-RCNN [11] to detect the human bounding box(es) and apply simple linear iterative clustering (SLIC) within every bounding box to partition the pixels into

N superpixels. $\{L_p^n | n = 1, ..., N\}$, refers to labeling superpixel n as body part p. For the rest of the paper, all analysis is carried within a specific bounding box although it can be easily extended to multiple bounding boxes.

We adopt the random variable formulation as [12] by using three binary random variables (x, y, z) in domains $x \in \{0,1\}^{N\times C}$, $y \in \{0,1\}^{\binom{N}{2}}$ and $z \in \{0,1\}^{\binom{N}{2}\times C^2}$. We use x_{nc} to show the superpixel's label and $y_{nn'}$ to describe the relationship of two superpixels n and n':

$$x_{nc} = \begin{cases} 1 & n \text{ is of class } c \\ 0 & \text{otherwise} \end{cases}, \quad y_{nn'} = \begin{cases} 1 & n, n' \text{ is of same person} \\ 0 & \text{otherwise} \end{cases} \tag{1}$$

Moreover, we use $z_{nn'cc'} = x_{nc}x_{n'c'}y_{nn'}$ to correlate x and y, where $z_{nn'cc'} = 1$ refers that superpixel n and n' belong to the same person (i.e., $y_{nn'} = 1$) and have labels c, c' respectively. Otherwise, $z_{nn'cc'} = 0$.Under this notion, we can easily write down the attributes of a desirable labeling scheme. First, we enforce each superpixel n to have precisely one label, either body part or background as: for each $n \in N$: $\sum_{c\in C} x_{nc} = 1$.

Therefore, if a superpixel pair n-n' belong to same person ($y_{nn'=1}$), n-n' should be labeled as one of the body parts rather than the background, i.e., for each $n, n' \in \binom{N}{2}$: $y_{nn'} \leq \sum_{c\in C, c\neq bg} x_{nc}, \quad y_{nn'} \leq \sum_{c\in C, c\neq bg} x_{n'c}$.

Further, the transitive closure property should hold: if superpixel pairs n-n' and n-n'' belong to the same person, then pair n' and n'' should also belong to the same person, i.e., for each $n, n', n'' \in \binom{N}{3} : y_{nn'} + y_{n'n''} - 1 \leq y_{nn''}$

Finally, for any superpixel pair n-n' and any $cc' \in C^2$ where $z_{nn'} = x_{nc}x_{n'c'}y_{nn'}$, for each $n, n' \in \binom{N}{2}$, we have: $x_{nc} + x_{n'c'} + y_{nn'} - 2 \leq z_{nn'cc'}$.

We refer to the four attributes as the *labeling constraints*.

3.2. Objective Function

Next, we put together an object function that obeys the labeling constraints above. Specifically, we model the problem as a refinement process: we obtain an initial segmentation score map P_s using state-of-art algorithm [2] along with an estimated pose L_j using [1]. We define:

$$\arg\min_{(x,z)\in\mathscr{D}} \mathscr{F} = \sum_{n=1}^{N}\sum_{c=1}^{C} \phi_{data}(n, c, P_s) + \sum_{nn'\in\binom{N}{2}} \sum_{c,c'} \phi_{smooth}(n, c, n', c', P_s, L_j) \tag{2}$$

where

$$\phi_{data}(n, c, P_s) = log\frac{1 - p_{nc}}{p_{nc}} \cdot x_{nc} \tag{3}$$

and

$$\phi_{smooth}(n, c, n', c', P_s, L_j) = log\frac{1 - p_{nn'cc'}(P_s, L_j)}{p_{nn'cc'}(P_s, L_j)} \cdot z_{nn'cc'}, \tag{4}$$

where $p_{nc} \in (0, 1)$ is the probability of superpixel n being labeled as class c , which can be extracted from the score map P_s stored as a $N \times C$ matrix.

To solve for the object function $\mathscr{F}$, we adopt the Fully-connected Conditional Random Field or FCRF approach. A dense FCRF consists of a data term and a smooth term, where the data term ϕ_{data} computes the unary potentials of superpixel patch n being labeled as c for every pair $(n, c) \in N \times C$. To compute the data term and correlate to our labeling constraints described in last section, we transform the original $\phi_{data}(n, n', P_s)$ to $log\frac{1-p_{nc}}{p_{nc}} \cdot x_{nc}$.

The smoothness term ϕ_{smooth} computes pairwise potentials of n and n' being labeled as c and c' respectively over all possible superpixel $nn' \in \binom{N}{2}$ and label pairs $cc' \in C^2$. Computing the smooth term is more complicated. Similar to the data term, we transform the original $\phi_{smooth}(n, c, n', c', P_s, L_j)$ into $log\frac{1-p_{nn'cc'}}{p_{nn'cc'}} \cdot z_{nn'cc'}$ where $p_{nn'cc'}$ is the probability of superpixels n and n' being from the same person and labeled as class c and c' respectively. In the following section, we discuss how to compute Pairwise Probability $p_{nn'cc'}$ from the training data.

4. COMPUTE PAIRWISE PROBABILITY

Recall that the pairwise term $p_{nn'cc'}$ is affected by both the joint location L_j (from the estimated pose) and the segmentation score map P_s. Based on above observation, we aim to train a mapping between each set of $\{n, n', c, c', P_s, L_j\}$ and $p_{nn'cc'}$, where $nn' \in \binom{N}{2}$, $cc' \in C^2$.

We design a feature descriptor $f(n, n', c, c', P_s, L_j)$ (for simplicity, we use f for the rest of the section) as a representation for each $\{n, n', c, c', P_s, L_j\}$ to train number of C^2 classifiers. Each trained classifier outputs $p_{nn'cc'}$ as the probability of $z_{nn'cc'} = 1$. In our implementation, we adopt a 11-d feature vector based on the estimated pose as:

$$f = [f_s(n, c, P_s, L_j)\, f_s(n', c', P_s, L_j)\, f_p(n, n', c, c', P_s, L_j)] \tag{5}$$

f consists of three components: $f_s(n, c, P_s, L_j)$ and $f_s(n', c', P_s, L_j)$ are the unary features computed for matching super-pixel patch n and n' against label c and c' respectively; the third component $f_p(n, n', c, c', P_s, L_j)$ corresponds to pairwise "similarity" between n and n' with respect to their labels. The first two terms are 5d whereas the last term is 1D, hence forming a 11d feature vector.

4.1. Unary Feature

Let's first consider the unary feature $f_s(n, c, P_s, L_j)$ computed as:

$$f_s = [p_{nc}, IoU_n^c, \rho_u(\mathscr{L}^n_{\kappa_c^1}, n), \rho_u(\mathscr{L}^n_{\kappa_c^2}, n), \delta_{ratio}(\mathscr{L}^{n_{cen}}_{O^{1}_{\kappa_c}\kappa^2_c})] \tag{6}$$

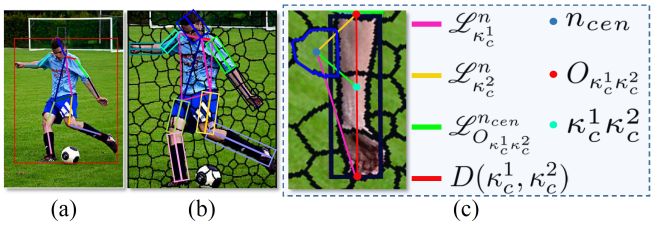

Fig. 3. Unary feature illustration. (a) is the human pose estimated by [1]. (b) shows we expand skeletons into rectangles to effectively framed each body parts. (c) depicts the line we defined in Sec. 4.1

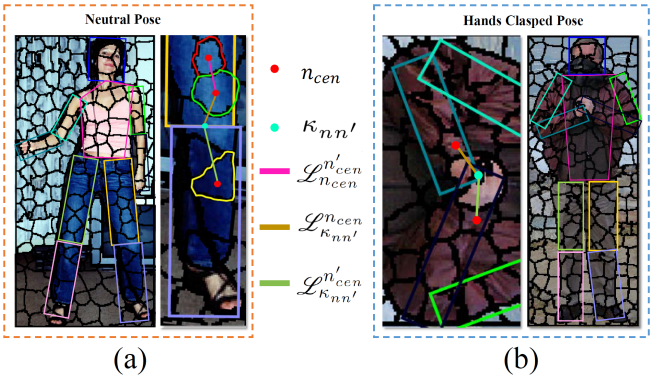

Fig. 4. Pairwise feature illustration. (a) shows a natural human pose: upper leg is adjacent to lower leg. While (b) shows variability of poses add different connectivities: clasped hands.

In the 5D feature vector, the first term computes the probability p_{nc} that super-pixel patch n labeled as c under the initial score map P_s. The second term augments each segment (a line segment between a pair of joints) in the estimated pose into a rectangle rec_c as shown in Fig. 3 where we use the canonical model to determine the width and height of the rectangle [13]. It then measures the percentage of the superpixel lying inside the bounding rectangle as $IoU^c_n \in (0, 1)$.

For the third and the fourth terms, recall each body part c should be associated with two joint (κ^1_c and κ^2_c) or a "bone" (line segment) $\kappa^1_c - \kappa^2_c$. We therefore connect the centroid n_{cen} of super-pixel n with the two joints to form two line segments $\mathscr{L}^{n}_{\kappa^1_c}$ and $\mathscr{L}^{n}_{\kappa^2_c}$ and we individually measure the percentage of the two segments that lie inside rec_c.

The final term measures the "distance" between the centroid of the superpixel to the "bone", i.e., by computing the distance from the centroid n_{cen} of the superpixel to the midpoint O of line segment $\kappa^1_c - \kappa^2_c$. Conceptually, the smaller the distance the more likely n should be labed as c. We normalize this distance by the length of the bone as δ_{ratio}.

Since each superpixel-label pair maps to a 5d feature vector, two superpixels and their potential labels map to a 10d vector.

4.2. Pairwise Feature

The last term $f_p(n, n', c, c', P_s, L_j)$ is computed as:

$$f_p = \begin{cases} \rho_{c=c'}(\mathscr{L}^{n'_{cen}}_{n_{cen}}, rec_c) \text{ when } c = c' \\ \dfrac{\rho_{c\neq c'}(\mathscr{L}^{n'_{cen}}_{\kappa^1_{c'}}, rec_{c'}) + \rho_{c\neq c'}(\mathscr{L}^{n_{cen}}_{\kappa^2_c}, rec_c)}{2} \text{ when } c \neq c' \end{cases} \tag{7}$$

We consider two cases: 1) the two superpixels labeled as the same body part, i.e., $c = c'$ or 2) they are labeled as different body parts, i.e., $c \neq c'$. For the former (Fig. 4 we connect the corresponding centroids of the two superpixels into a line segment $\mathscr{L}^{n'_{cen}}_{n_{cen}}$ and compute the percentage of pixels that lie in rec_c as $\rho_{c=c'}(\mathscr{L}^{n'_{cen}}_{n_{cen}}, rec_c)$.

For the latter, if the two body parts c and c' are adjacent to each other, we simply deem such labeling is in-feasible and set the value to be 0. For example, the head should be adjacent to the torso, the torso to the upper arm, etc. In addition, various movements can cause different connectivities, as shown in Fig. 4(b). This can be determined by the movement of the body, e.g, crossing arms, arms on legs, etc.

In the adjacency case, every pair of body part c and c' should have a common joint $\kappa_{cc'}$. We therefore can connect the common joint with the centroid of the two superpixels as $\mathscr{L}^{n_{cen}}_{\kappa_{cc'}}$ and $\mathscr{L}^{n'_{cen}}_{\kappa_{cc'}}$. We then compute the percentage of the line segment $\mathscr{L}^{n_{cen}}_{\kappa_{cc'}}$ lying within rec_c as $\rho_{c\neq c'}(\mathscr{L}^{n_{cen}}_{\kappa_{cc'}}, rec_c)$. Similarly, we can compute $\rho_{c\neq c'}(\mathscr{L}^{n'_{cen}}_{\kappa_{cc'}}, rec_{c'})$ accordingly. We use the average of the two values as the last term in the feature vector.

4.3. Training and Optimization

Once we define the f as described above, we set out to train our classifiers $\{w_{cc'}, cc' \in C^2\}$ using non-linear logistic regression. Specifically, we use the VOC12 human part training dataset that contains 1,712 images. After the training process, we can compute the pairwise superpixel probability $p_{nn'cc'}(P_s, L_j)$ as:

$$\begin{aligned} p(z_{nn'cc'} = 1) &= \frac{1}{1 + \exp(-w_{cc'} \cdot f(n, n', c, c', P_s, P_j))} \\ &= \langle w_{cc'}, f(n, n', c, c', P_s, P_j) \rangle \end{aligned} \tag{8}$$

Finally, we solve Eq. 2 using the state-of-the-art integer linear programming (ILP) solver Guruobi. For initialization, we simply use the semantic segmentation results of DeepLab where every $y_{nn'}$ is initialized by detecting if the two patches lie in the same human bounding box. If a patch lies in the intersection regions of multiple bounding boxes, we assign it to the closest bounding box. We then compute $z_{nn'cc'} = x_{nc}x_{n'c'}y_{nn'}$.

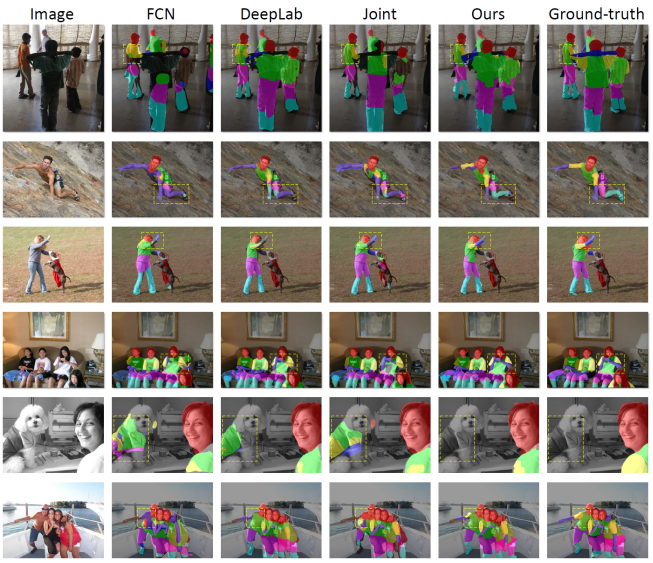

Fig. 5. Our semantic segmentation results on VOC PASCAL-Person-Part. We compare our method with recent state-of-the-art methods FCN [4], DeepLab(our own implementation) [2] and Xia et al. [3].

Next, we find a feasible solution of set $\mathscr{D}$ by only considering the data term. Finally, we add the smoothness term and iteratively refine $\mathscr{D}$ under the pairwise constraints (Sec. 3.1). The iterative stops when the change is lower than a pre-set threshold (1% in our case).

5. EXPERIMENTS

We first evaluate our technique on publicly available human part datasets, VOC PASCAL-Person-Part [9]. They provide semantic part segmentation of human from the PASCAL VOC2010 dataset. We use 7 semantic labels: head, torso, upper/lower arm, upper/lower leg in our experiment. For training and validation, we use $1,716$ training images and $1,817$ testing images. We compare our method with the state-of-art algorithms, i.e., FCN [4], DeepLab [2] and the work by Xia et al. [3]. All experiments (including training and testing) were performed off-line on a PC with CPU Intel Core i7-5820K, 32GB memory, and NVIDIA TITAN X GPU. On computational time, ILP optimization described in Sec. 3.2 on average takes about 1 minute on images of a resolution 335×500.

Initial pose estimation and part segmentation. Our method takes pose estimation as input and outputs human part segmentation. To obtain initial human pose estimations, we employ two pre-trained frameworks, e.g., AlphaPose [1] for pose extraction and Mask RCNN [11] for human bounding box prediction. For initial human part segmentation, we adopt an end-to-end ResNet-based learning architecture DeepLab [2]. After initialization, we obtain, for each image in the dataset, the bounding box based pose results along with a human part score map with 7 classes: 6 human parts and 1 background label.

Human part semantic segmentation evaluation. We conduct our experiments on 1,817 testing images from the VOC PASCAL-Person-Part dataset. Fig. 5 shows visual comparisons of our technique vs. the recent semantic segmentation solutions by Long et al. [4], Chen et.al [2] and Xia et.al [3]. We refer the reviewers to the supplementary materials for the complete results.

Method	Size XS	Size S	Size M	Size L
FCN [4]	31.2	44.4	50.1	49.5
DeepLab [2]	29.3	45.4	52.7	54.3
Xia et al. [3]	39.8	50.2	54.1	54.5
Attention [5]	37.6	49.8	55.1	55.5
Our	**40.0**	**51.3**	**57.1**	**58.3**

Table 1. Mean IOU(mIOU) of human part segmentation on four different sizes of human. We observe that our Pose2Body technique achieves more accurate labeling results over all sizes.

We observe that our Pose2Body technique achieves much better visual quality on scenes that exhibit heavy occlusions and cluttered environments. For example, in the climber image, our method manages to correctly label the lower legs whereas previous methods fail as shown in the second row of Fig. 5. This is because occlusions impose significant challenges to directly label the body parts: the arm and leg were deemed to be a single part in previous approaches whereas our technique successfully separates their corresponding skeletons and hence their labels. In the dog-human image, state-of-the-art deem the dog be a part of the human and hence fails to correctly label the body parts. Our technique, in contrast, manages to determine the dog should not be a part of the skeleton (due to violations of continuity) and correctly label the human and her body parts.

For quantitative evaluation, we compute pixel-wise accuracy measurement, i.e., mean pixel intersection-over-union(mIOU) over each class. In Table 2, we compare our method with four prior art [5, 2, 4, 3]. Our method outperforms these solutions by approximately 2% on average over all parts. In particular, our technique achieves much higher mIOU on Torso, L-arms, U-legs, L-legs and Background.

Method	Head	Torso	U-arms	L-arms	U-legs	L-legs	Background	Ave.
FCN [4]	77.99	52.84	37.75	36.67	32.66	30.47	92.39	51.54
Attention [5]	**81.47**	59.06	44.15	42.50	38.28	35.62	93.65	56.39
Xia et al. w/o pose [3]	79.83	59.72	43.84	40.84	40.49	37.23	93.55	56.50
Xia et al. [3]	80.21	61.36	47.53	43.94	41.77	38.00	93.64	58.06
DeepLab [2]	80.12	61.47	**47.62**	43.53	41.65	37.00	93.59	57.86
Our	81.27	**61.69**	45.79	**46.88**	**45.04**	**43.54**	**94.49**	**59.82**

Table 2. Mean IOU(mIOU) of human semantic part segmentation on PASCAL human part dataset with 6 body parts. We show that our technique outperform the stat-of-the-art methods not only on average over all parts, but also achieves much higher mIOU values in most individual human parts.

This is because pose estimation manages to restrict the label region within the bounding boxes so that the leg and arm regions become better separated and hence our computed features become more separated. [4, 2] produce higher errors on the upper & lower leg parts. Recall that Xia et al. [3] also refine the initialized results used two types of initializations: the first using VGG-16 (boosting from 56.50 to 58.06) and the second using ResNet-101 (boosting from 62.66 to 64.39), achieving about 1.56 and 1.73 gain. We use initialization obtained from DeepLab and we manage to boost the performance from 57.86 to 60.02, achieving about 2.16 gain.

Next, we follow [10] and conduct experiments on variable human sizes. We categorize all human instances detected by the human detector into four different sizes in terms of the bounding box size S_b and calculate the average mIOU on individual categories. Specifically, we use four sizes: size XS where $S_b \in [0, 80]$, size S where $S_b \in [80, 140]$, size M where $S_b \in [140, 220]$, and size L where $S_b \in [220, 520]$. Table. 1 shows the final results. We show our method outperforms the baseline techniques on all sizes except size XS.

6. CONCLUSIONS AND DISCUSSIONS

We have presented a novel human parts segmentation scheme based on human pose estimation. Our Pose2Body technique partitions pixels within the bounding box of each human subject into superpixels and employs the conditional random field (CRF) model for labeling each superpixels. Specifically, we have designed a class of features that correlate the appearance and location of superpixels with the estimated pose as well as correlate pairs of superpixels. We have developed reliable solutions to solve for the CRF using integer linear programming and have conducted comprehensive experiments to demonstrate the advantages of our technique over the state-of-the-art. A limitation of our approach is that it still relies on the estimated pose as input rather than simultaneously conducting pose estimation and semantic labeling. In the future, we intend to explore uniformly modeling both problems in terms of label assignment.

7. REFERENCES

[1] Haoshu Fang, Shuqin Xie, Yu-Wing Tai, and Cewu Lu, "Rmpe: Regional multi-person pose estimation," *2017 IEEE International Conference on Computer Vision (ICCV)*, pp. 2353–2362, 2017.

[2] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L. Yuille, "Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs," *IEEE transactions on pattern analysis and machine intelligence*, 2017.

[3] Fangting Xia, Peng Wang, Xianjie Chen, and Alan L. Yuille, "Joint multi-person pose estimation and semantic part segmentation," *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6080–6089, 2017.

[4] Evan Shelhamer, Jonathan Long, and Trevor Darrell, "Fully convolutional networks for semantic segmentation," *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3431–3440, 2015.

[5] Liang-Chieh Chen, Yi Yang, Jiang Wang, Wei Xu, and Alan L. Yuille, "Attention to scale: Scale-aware semantic image segmentation," *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3640–3649, 2016.

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

[7] Fangting Xia, Jun Zhu, Peng Wang, and Alan L. Yuille, "Pose-guided human parsing by an and/or graph using pose-context features," in *AAAI*, 2016.

[8] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L. Yuille, "Semantic image segmentation with deep convolutional nets and fully connected crfs," *CoRR*, vol. abs/1412.7062, 2014.

[9] Peng Wang, Xiaohui Shen, Zhe L. Lin, Scott Cohen, Brian L. Price, and Alan L. Yuille, "Joint object and part segmentation using deep learned potentials," *2015 IEEE International Conference on Computer Vision (ICCV)*, pp. 1573–1581, 2015.

[10] Fangting Xia, Peng Wang, Liang-Chieh Chen, and Alan L. Yuille, "Zoom better to see clearer: Human and object parsing with hierarchical auto-zoom net," in *ECCV*, 2016.

[11] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross B. Girshick, "Mask r-cnn," *2017 IEEE International Conference on Computer Vision (ICCV)*, pp. 2980–2988, 2017.

[12] Leonid Pishchulin, Eldar Insafutdinov, Siyu Tang, Bjoern Andres, Mykhaylo Andriluka, Peter V. Gehler, and Bernt Schiele, "Deepcut: Joint subset partition and labeling for multi person pose estimation," *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4929–4937, 2016.

[13] Leonid Sigal and Michael J Black, "Predicting 3d people from 2d pictures," in *International Conference on Articulated Motion and Deformable Objects*. Springer, 2006, pp. 185–195.