# Joint Multi-Person Pose Estimation and Semantic Part Segmentation


Fangting Xia*, Peng Wang*, Xianjie Chen*, Alan Yuille+

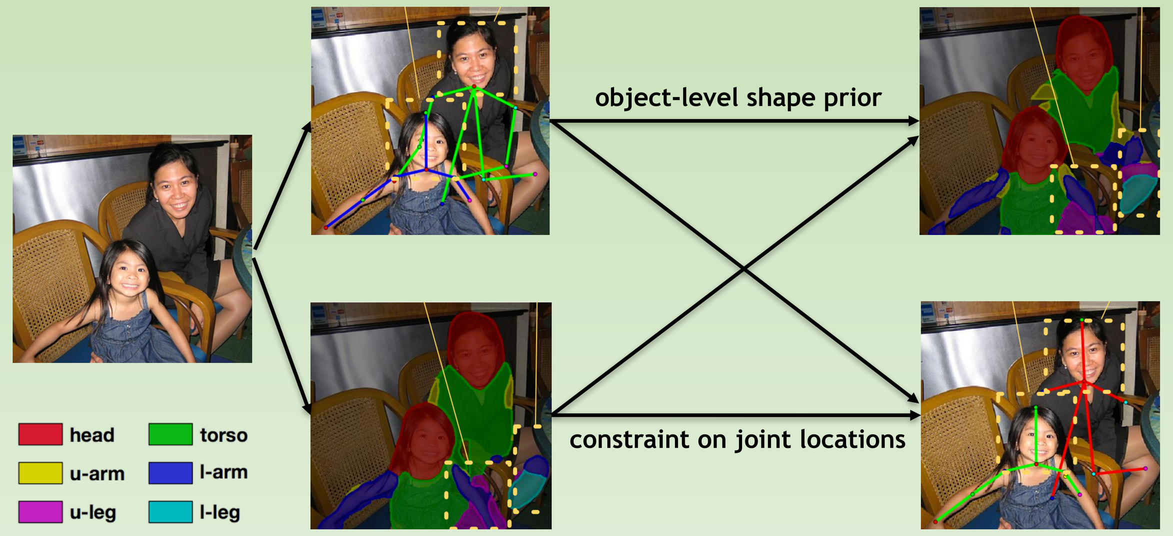
*University of California, Los Angeles +Johns Hopkins University



IEEE 2017 Conference on Computer Vision and Pattern Recognition

CVPR July 21-26 Honolulu 2017


## Motivation




Human pose estimation and semantic part segmentation are two complementary tasks. Can we use the more available pose annotations to help semantic part segmentation?

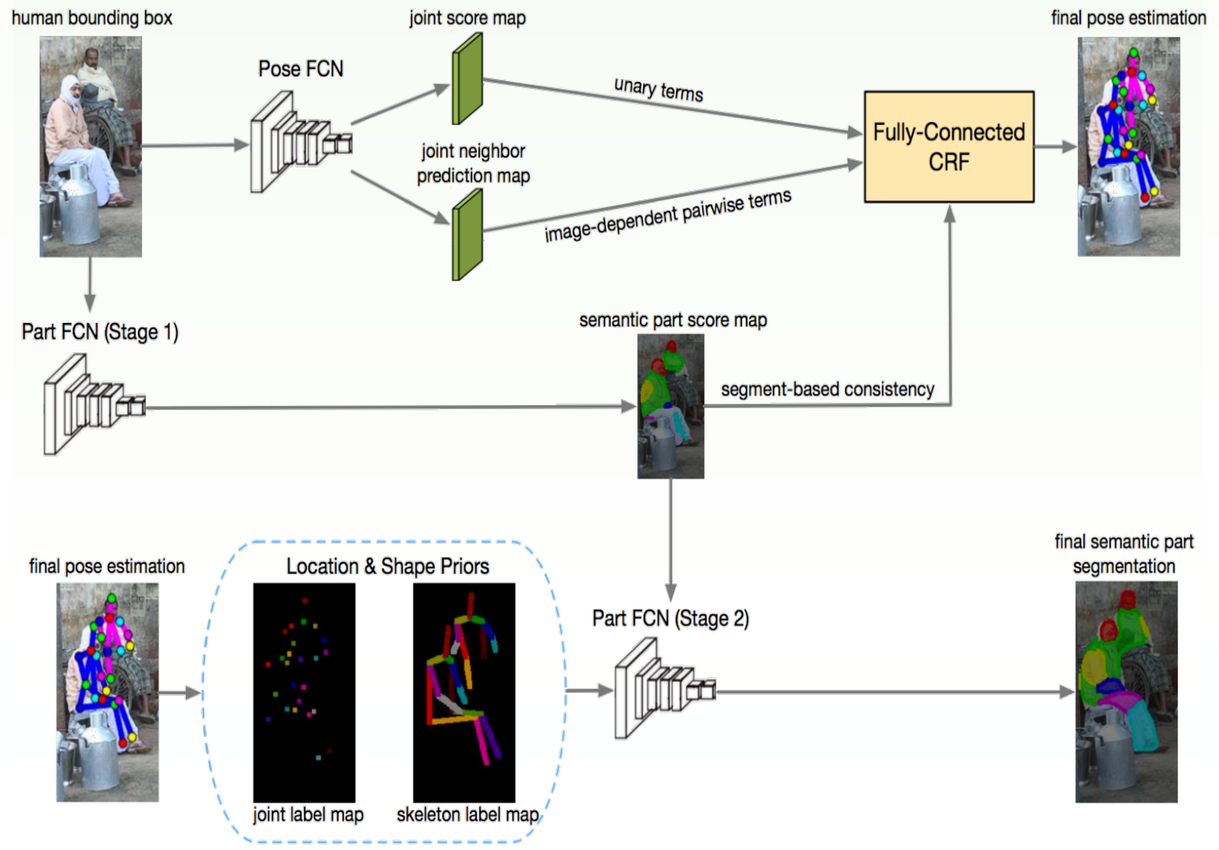
## Introduction

### ❖ Dataset

- ❑ Extends the challenging PASCAL-Person-Part dataset[1] with pose joint annotations and make the annotations public[2].

### ❖ Model Highlights

- ❑ Deep-based iterative framework for joint estimation of human pose and semantic parts.
  - segment-joint smoothness term to force consistency between joint locations and given semantic part masks
  - pose-based regularization cues for part segmentation
- ❑ "Auto-zoom" strategy to handle large scale variation.
- ❑ Surpasses competing methods by a large margin.
  - pose estimation: 10.6% mAP with faster inference
  - semantic part segmentation: 1.5% mIOU overall and 5% for small-scale people




## Our Approach

### ❖ Overview

- ❑ Extracts human bounding boxes using Faster R-CNN[3].
- ❑ Resizes image regions in bounding boxes using "auto-zoom"[4] so that small people are enlarged and extra large people are shrunk to a fixed size.
- ❑ Resized image regions serve as input to Pose FCN and Part FCN, which output initial estimation of pose joint potential and semantic part potential respectively.
  - Pose FCN: uses architecture ResNet-101[5]; outputs unary joint score map and pairwise joint neighbor location regression map
  - Part FCN: uses architecture DeepLab-LargeFOV[6]; outputs initial segment part score map
- ❑ Refines pose estimation using semantic part potential.
  - generates joint proposals from unary joint score map
  - computes segment-joint smoothness terms for joint proposals based on the initial segment part potential
  - selects and assembles joint proposals using a FCRF
- ❑ Refines semantic part potential through Part FCN with location and shape priors inferred from pose estimation.
- ❑ For both tasks, merge the refined result of each bounding box as the final result for the image.

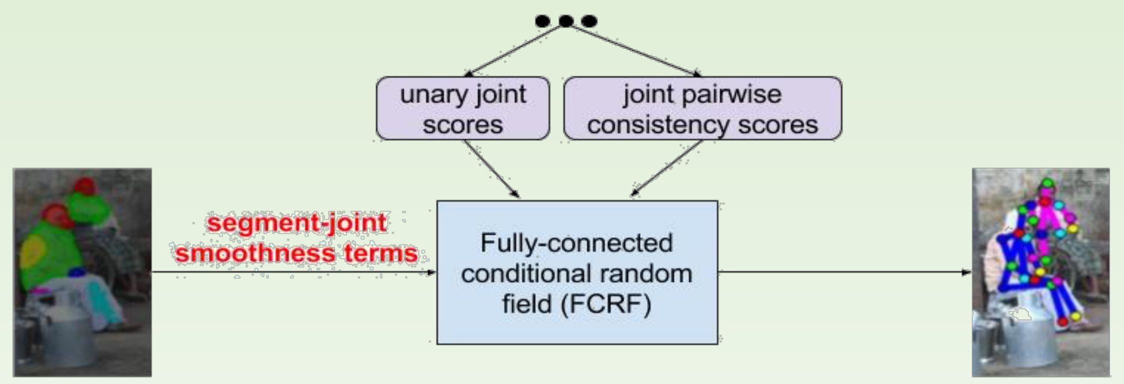
### ❖ Pose Estimation Component

- ❑ Fully-connected CRF ( $\mathcal{G} = \{\mathcal{V}, \mathcal{E}\}$ )
  - $\mathcal{V} = \{c_1, c_2, \ldots, c_n\}$ represents all candidate locations for joints; $\mathcal{E} = \{(c_i, c_j) | i = 1, 2, \ldots, n, j = 1, 2, \ldots, n, i < j\}$
  - predict joint type $l_{c_i} \in \{0, \cdots, K\}$ and whether two joints belong to the same person $l_{c_i,c_j} \in \{0, 1\}$

$$\mathcal{L} = \{l_{c_i} | c_i \in \mathcal{V}\} \cup \{l_{c_i,c_j} | (c_i, c_j) \in \mathcal{E}\}$$

  - target to optimize:

$$\min_{\mathcal{L}} \sum_{c_i \in \mathcal{V}} \psi_i(l_{c_i}) + \sum_{(c_i,c_j) \in \mathcal{E}} \psi_{i,j}(l_{c_i}, l_{c_j}, l_{c_i,c_j})$$




Pose estimation: FCRF with our novel segment-joint smooth terms for instance clustering and joint labeling.

### ❖ Pose Estimation Component (Cont.)

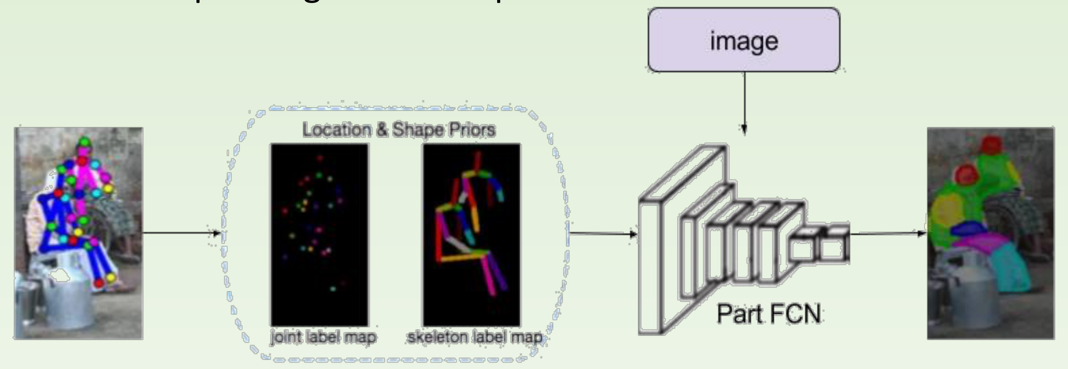
- ❑ Segment-joint smoothness term

$$\mathbf{f}(c_i, c_j, l_{c_i}, l_{c_j} | \mathbf{P}_s) = [\mathbf{f}_u(c_i, l_{c_i} | \mathbf{P}_s) \quad \mathbf{f}_u(c_j, l_{c_j} | \mathbf{P}_s) \quad \mathbf{f}_p(c_i, c_j, l_{c_i}, l_{c_j} | \mathbf{P}_s)]$$

  - represents the compatibility between joints $c_i, c_j$ and part segmentation score map $\mathbf{P}_s$
  - $\mathbf{f}_u$: the relative position of joint wrt. its associated part(s)
  - $\mathbf{f}_p$: the overlap between line $<c_i, c_j>$ and its associated part

### ❖ Semantic Part Segmentation Component

- ❑ Pose joint label map and skeleton label map as additional feature maps to the original part segmentation feature maps for Part FCN.
- ❑ Stacks 3 additional conv layers (7*7*128) + Relu layer to produce the final part segmentation potential.




Semantic part segmentation: two-stream Part FCN with our novel location and shape priors inferred from pose.

### ❖ Evaluation of Human Pose Estimation

- ❑ Extensive experiments on PASCAL-Person-Part[1,2].
  - contains large variation of human pose and scale
  - 14 annotated joint types and 6 semantic part types
  - 1716 training images and 1817 images for testing
- ❑ Competing methods
  - Chen & Yuille[7]: tree-structured model for single-person pose estimation in presence of occlusion, using DCNN features
  - Deeper-Cut[8]: graphical model that jointly performs multi-person pose estimation
  - AOG-Simple: an And-Or graph with only unary joint scores and pairwise geometric constraint between neighboring joints
  - AOG-Seg: an And-Or graph that adds to AOG-Simple segment-joint smoothness terms

| Method | Head | Shoulder | Elbow | Wrist | Hip | Knee | Ankle | U-Body | Total (mAP) |
|---|---|---|---|---|---|---|---|---|---|
| Chen & Yuille | 45.3 | 34.6 | 24.8 | 21.7 | 9.8 | 8.6 | 7.7 | 31.6 | 21.8 |
| Deeper-Cut | 41.5 | 39.3 | 34.0 | 27.5 | 16.3 | 21.3 | 20.6 | 35.5 | 28.6 |
| AOG-Simple | 56.8 | 29.6 | 14.9 | 11.9 | 6.6 | 7.3 | 8.6 | 28.3 | 19.4 |
| AOG-Seg | **58.5** | 33.7 | 17.6 | 13.4 | 7.3 | 8.3 | 9.2 | 30.8 | 21.2 |
| Our Model (w/o seg) | 56.8 | 52.1 | 42.7 | 36.7 | 21.9 | 30.5 | 30.4 | 47.1 | 38.7 |
| Our Model (final) | 58.0 | **52.1** | **43.1** | **37.2** | **22.1** | **30.8** | **31.1** | **47.6** | **39.2** |

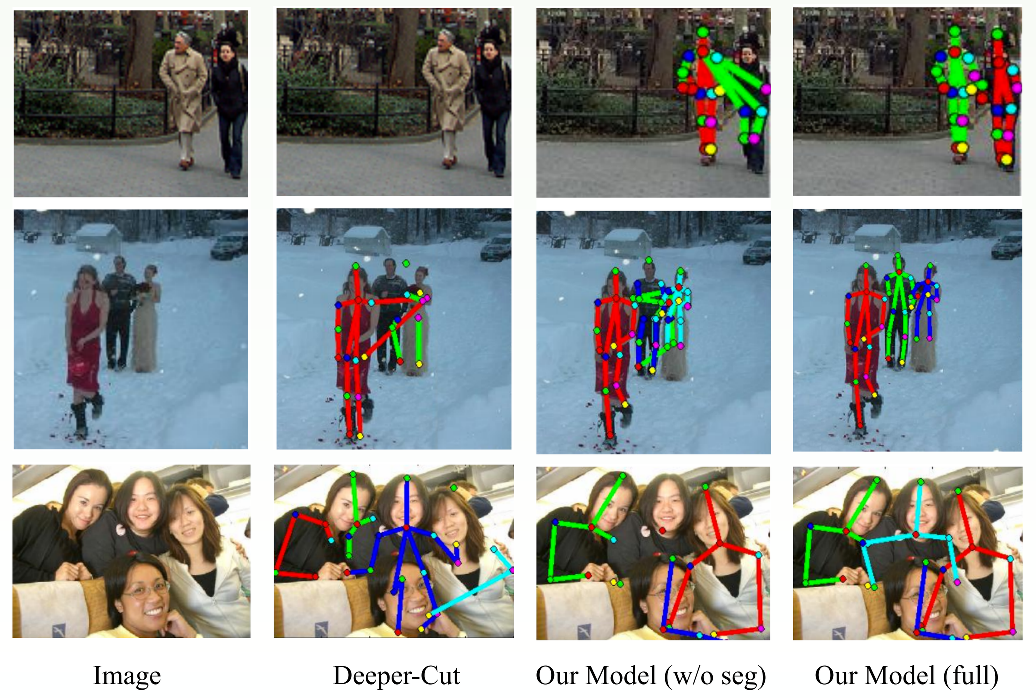
Mean average precision (mAP) (%) of pose estimation on PASCAL-Person-Part.

## Experimental Results

### ❖ Evaluation of Human Pose Estimation (Cont.)

| Method | Forehead | Neck | Shoulder | Elbow | Wrist | Hip | Knee | Ankle | Ave. |
|---|---|---|---|---|---|---|---|---|---|
| Chen & Yuille | 37.5 | 29.7 | 51.6 | 65.9 | 72.0 | 70.5 | 79.9 | 78.6 | 60.7 |
| Deeper-Cut | 32.1 | 30.9 | 37.5 | 44.6 | 53.5 | 53.9 | 65.8 | 67.8 | 48.3 |
| AOG-Simple | 33.0 | 33.2 | 66.7 | 82.3 | 90.5 | 89.7 | 101.3 | 101.1 | 74.7 |
| AOG-Seg | 32.2 | 31.6 | 59.8 | 72.4 | 85.1 | 85.7 | 97.1 | 92.7 | 69.6 |
| Our Model (w/o seg) | 27.7 | 26.9 | 33.1 | 40.2 | 47.3 | 51.8 | 54.6 | 53.4 | 41.9 |
| Our Model (final) | **26.9** | **26.1** | **32.7** | **39.5** | **45.3** | **50.9** | **52.3** | **51.8** | **40.7** |

Average distance of keypoints (ADK) (%) of pose estimation on PASCAL-Person-Part.



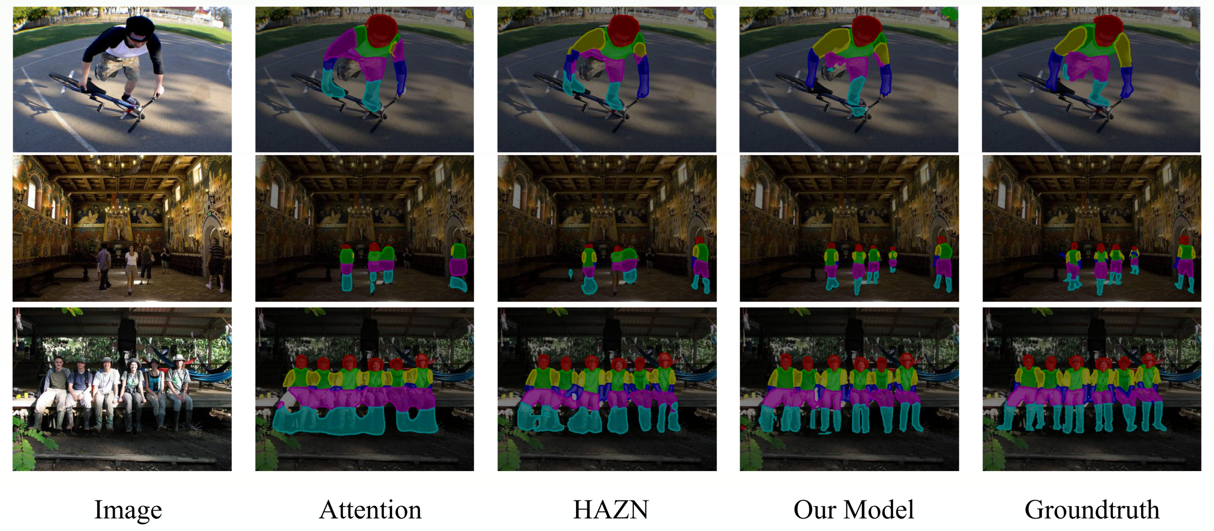

Visual comparison of pose estimation on PASCAL-Person-Part. Our model gives better prediction of heads, arms and legs, and is especially better at handling people of small scale.

### ❖ Evaluation of Human Semantic Part Segmentation

- ❑ Competing methods
  - Attention[9]: FCN-based model with scale attention network
  - HAZN[4]: hierarchical model that adapts to object/part scales

| Method | Head | Torso | U-arms | L-arms | U-legs | L-legs | Background | Ave. |
|---|---|---|---|---|---|---|---|---|
| Attention [5] | 81.47 | 59.06 | 44.15 | 42.50 | 38.28 | 35.62 | 93.65 | 56.39 |
| HAZN [33] | 80.76 | 60.50 | 45.65 | 43.11 | 41.21 | 37.74 | 93.78 | 57.54 |
| Our model (VGG-16, w/o pose) | 79.83 | 59.72 | 43.84 | 40.84 | 40.49 | 37.23 | 93.55 | 56.50 |
| Our model (VGG-16, final) | 80.21 | 61.36 | 47.53 | 43.94 | 41.77 | 38.00 | 93.64 | 58.06 |
| Our model (ResNet-101, w/o pose) | 84.95 | 67.21 | 52.81 | 51.37 | 46.27 | 41.03 | 94.96 | 62.66 |
| Our model (ResNet-101, final) | **85.50** | **67.87** | **54.72** | **54.30** | **48.25** | **44.76** | **95.32** | **64.39** |

Mean pixel IOU (mIOU) (%) of semantic part segmentation on PASCAL-Person-Part.




Visual comparison of semantic part segmentation on PASCAL-Person-Part. Our model estimates the overall configuration more accurately and gives clearer details of arms and legs, especially for small-scale people.

## Experimental Results (Cont.)

### ❖ Evaluation of Human Semantic Part Segmentation

| Method | Size XS | Size S | Size M | Size L |
|---|---|---|---|---|
| Attention [5] | 37.6 | 49.8 | 55.1 | 55.5 |
| HAZN [33] | 47.1 | 55.3 | 56.8 | 56.0 |
| Our model (ResNet-101, w/o pose) | 40.4 | 54.4 | 60.5 | 62.1 |
| Our model (ResNet-101, final) | **53.4** | **60.9** | **63.0** | **62.8** |

Mean pixel IOU (mIOU) (%) of semantic part segmentation wrt. size of human instance on PASCAL-Person-Part. Our model improves the results by over 5% for small-scale people.

## Conclusions

- ❖ We present an efficient framework that demonstrates the complementary property of human pose estimation and semantic part segmentation in natural multi-person images.
- ❖ For pose estimation, we adopt a FCRF using deep-learned joint scores and part segment-based consistency features, giving more accurate localization of joints, especially for arm joints and leg joints.
- ❖ For semantic part segmentation, we train a two-stream FCN that uses estimated pose configurations as shape and location priors, successfully correcting pose errors and giving clearer details of arms and legs.
- ❖ We also adopt an effective "auto-zoom" strategy that deals with object scale variation for both tasks and speeds up the inference of FCRF by 40 times.

## References


[1] X. Chen et al. Detect what you can: Detecting and representing objects using holistic models and body parts. In CVPR, 2014.

[2] https://sukixia.github.io/paper.html

[3] S. Ren et al. Faster R-CNN: Towards real-time object detection with region proposal networks. In NIPS, 2015.

[4] F. Xia et al. Zoom better to see clearer: Human and object parsing with hierarchical auto-zoom net. In ECCV, 2016.

[5] K. He et al. Deep residual learning for image recognition. In CVPR, 2016.

[6] L. Chen et al. Semantic image segmentation with deep convolutional nets and fully connected CRFs. In ICLR, 2015.

[7] X. Chen at al. Parsing occluded people by flexible compositions. In CVPR, 2015.

[8] E. Insafutdinov et al. Deeper-Cut: A deeper, stronger, and faster multi-person pose estimation model. In ECCV, 2016.

[9] L. Chen et al. Attention to scale: Scale-aware semantic image segmentation. In CVPR, 2016.


## Acknowledgements


We are deeply grateful for the support from ONR N00014-15-1-2356, NSF award CCF-1317376 and Army Research Office ARO 62250-CS. Also, great thanks to NVIDIA for providing us with free GPUs for training deep neural networks.