

Weakly Supervised Body Part Segmentation with Pose based Part Priors

Zhengyuan Yang¹, Yuncheng Li², Linjie Yang³, Ning Zhang⁴, and Jiebo Luo¹

¹Department of Computer Science, University of Rochester

²Google Inc., ³ByteDance AI Lab, ⁴Vaitl Inc.

Abstract—Human body part segmentation refers to the task of predicting the semantic segmentation mask for each body part. Fully supervised body part segmentation methods achieve good performances but require an enormous amount of effort to annotate part masks for training. In contrast to high annotation costs needed for a limited number of part mask annotations, a large number of weak labels such as poses and full body masks already exist and contain relevant information. Motivated by the possibility of using existing weak labels, we propose the first weakly supervised body part segmentation framework. The core idea is first converting the sparse weak labels such as keypoints to the initial estimate of body part masks, and then iteratively refine the part mask predictions. We name the initial part masks estimated from poses the “part priors”. with sufficient extra weak labels, our weakly supervised framework achieves a comparable performance (62.0% mIoU) to the fully supervised method (63.6% mIoU) on the Pascal-Person-Part dataset. Furthermore, in the extended semi-supervised setting, the proposed framework outperforms the state-of-art methods. Moreover, we extend our proposed framework to other keypoint-supervised part segmentation tasks such as face parsing.

I. INTRODUCTION

Body part segmentation [1], [2] aims to decompose humans into semantic part regions with the predicted part masks. Body part segmentation helps various vision tasks such as action recognition [3], person re-id [4], and image generation [5]. Its potential applications include but are not limited to online shopping [6] and surveillance systems [7]. Previous studies on part segmentation [8], [9], [1], [2] mostly follow a fully supervised setting and thus require manually annotated part masks as supervision during training. However, the number of annotations for body part segmentation is often limited due to the high annotation costs of labeling the pixel-level part masks. The Pascal-Person-Part dataset [8], [1] is the largest body part segmentation dataset by far but contains only 3.5K images in total.

On the other hand, vast amounts of available weak annotations such as human poses and full body masks already exist and contain information related to the part segmentation task. In particular, human poses [10], [11], [12], [13] provide cues for high-level human structures, *i.e.*, the coarse spatial locations of the human part. The widely used pose estimation dataset [14] contains more than 30K images with 150K pose annotations. The full human mask is also widely available, as “person” is a semantic class in many semantic segmentation datasets [14]. In this study, we focus on exploiting such abundant existing weak supervisions for part segmentation. Our

motivation of utilizing existing weak supervisions is similar to weakly supervised semantic segmentation [15], [16]. However, directly applying weakly supervised semantic segmentation methods onto the part segmentation task generates poor results due to the more complicated scenes and labeling spaces. There are 17.9 instances per image on average in PASCAL-person-part compared to 2.8 object instances in PASCAL semantic segmentation.

In this study, we investigate using human poses and full masks as the alternative supervision for training the body part segmentation. The core idea is to convert the sparse human poses into an initial part mask estimates, namely the part priors, and iteratively refine the mask predictions. Specifically, the first step is to convert poses into initial part mask estimates by exploiting the prior knowledge of human structures, *e.g.*, pixels around the connection between joints “shoulder” and “elbow” are likely to be “upper arm.” The uncertain boundary regions are left blank initially and are gradually recovered by an iterative refinement module. The intermediate mask predictions during the iteration are refined by the Conditional Random Field (CRF) [17] and are used as the supervision to step by step improve the segmentation network. In this paper, we study how to generate and refine the part priors for weakly supervised part segmentation.

To the best of our knowledge, this is the first feasible solution on weakly supervised body part segmentation. On Pascal-Person-Part, our proposed framework achieves comparable performance (62.0% mIoU) to the fully supervised state-of-the-art method (63.6% mIoU) [18], with additional weak supervision used. When compared in a semi-supervised setting that part masks from Pascal-Person-Part are used in training, the framework outperforms the state-of-the-art semi-supervised methods [2]. We show that the widely available weak supervisions improve the performance of body part segmentation by a large margin. Furthermore, our part segmentation framework can be extended to other keypoint supervised segmentation tasks. We extend our weakly supervised framework onto the face parsing task on the Helen [19] and AFLW dataset [20].

How does the proposed framework obtain significant improvements, and why can weak labels help the part parsing task? We show empirically in Section IV-D that the successful exploitation of extra weak data, which has not been previously leveraged, is the main reason for the good performance. Qualitative results show that the gain mainly comes from

reducing the error in body structure prediction, *e.g.*, confusing arms with legs. Furthermore, extensively experiments show the superiority and robustness of the proposed pose based part prior. The new problem formulation, the significant performance improvements, and the effective mechanisms to achieve the goal, constitute the main contributions of this work.

II. RELATED WORKS

Human body part segmentation. Human body part segmentation [9], [1], [2] is the task of generating body part masks based on human structures. Xia *et al.* [1] propose to jointly conduct body part segmentation and pose estimation, and show that the two complementary tasks could help each other. Fang *et al.* [2] propose a data augmentation method for part segmentation based on pose similarities. Part segmentation is related to human parsing [21], [22], [23], [24], [25], [26], [27], as both tasks aim to predict pixel-level semantic masks for human parts. The major difference is that body part segmentation focuses on predicting part masks that directly reflect body structures, *e.g.*, “upper/ lower arms.” On the contrary, human parsing contains clothing and object classes such as “sunglasses,” “hats,” “coats,” *etc.*

Weakly supervised semantic segmentation. Our study is also related to weakly supervised semantic segmentation. Frequently used weak supervisions for semantic segmentation include scribbles [16], [28], bounding boxes [29], [30], [31], points [15] and image labels [32], [33], [34], [35]. Despite the promising results achieved on the semantic segmentation task, existing weakly supervised segmentation methods require the saliency of target regions in color space. Directly applying the previous methods to more complicated scenarios, such as scene parsing [36] and part segmentation, generates limited performance. To solve the weakly supervised body part segmentation task, we use the human structure knowledge with the proposed pose based part prior. Our study is also related to previous explorations on transforming human poses into human parts [10], [37].

III. METHODOLOGY

A. Overview

We first introduce the problem formulation of body part segmentation. Given an input RGB image $I_n \in \mathbb{R}^{H \times W \times 3}$ of size $H \times W$ in training set D with N samples, the objective of body part segmentation is to classify each pixel into one of the C body part classes or background as an output mask. In fully supervised approaches, pixel-level annotations $L_n \in \{0, 1, \dots, C\}^{H \times W}$ are manually labeled, and the network is trained with the per-pixel cross-entropy loss ℓ . Such methods show good performance on tasks with sufficient pixel-level annotations. However, the number of body part annotations is still limited due to the high labeling costs. In this study, we investigate the approach of training body part segmentation models with weak supervisions. Capitalizing on large scale datasets such as MS-COCO [14] that contain poses and full mask annotations, we train the body part segmentation framework *without* any pixel-level part annotations.



Fig. 1: Examples of pose based part priors on Pascal-Person-Part. The colored, black and white regions denote foreground part priors, background areas, and uncertain areas, respectively.

The overall structure of the iterative framework is shown in Figure 2. In order to learn dense masks with sparse keypoint coordinates, we convert the poses into part prior masks by drawing geometric shapes between the corresponding keypoint locations. The conversion rule is designed only to cover the most confident regions, *i.e.*, to generate part priors that have high precision. Instead of learning the detailed part shapes, the objective of the part prior is to provide cues of the coarse body part location. On the estimated boundary regions, the part prior label is left blank and is later gradually recovered with iterative refinement. We use the generated prior masks to train a segmentation network with the proposed structure loss and mask loss. To recover the uncertain regions and get more accurate mask predictions, we introduce the iterative refinement module. The initially predicted masks on the training set are refined by a CRF and are adopted as extra training labels for the next iteration. We show that the refined masks contain more shape details as CRF exploits the low-level image features. Since the CRF refinement might lead to degeneration in structure information, both part priors and the refined mask predictions are used as supervision in the iterative training. Self-paced learning is further adopted to select reliable enough predictions as supervision during iterative refinement. Furthermore, we find that end-to-end CRF refinement such as CRFasRNN [38] does not work well on this task due to the missing of full supervisions.

B. Pose based Part Priors

The human pose $P = \{(x_j, y_j)\}_{j=1}^J$ is a set of keypoints on the body structure. As a sparse representation of human body structures, poses are not directly compatible with traditional fully supervised segmentation methods, where per-pixel annotations are required. Previous studies [1], [3], [2] have

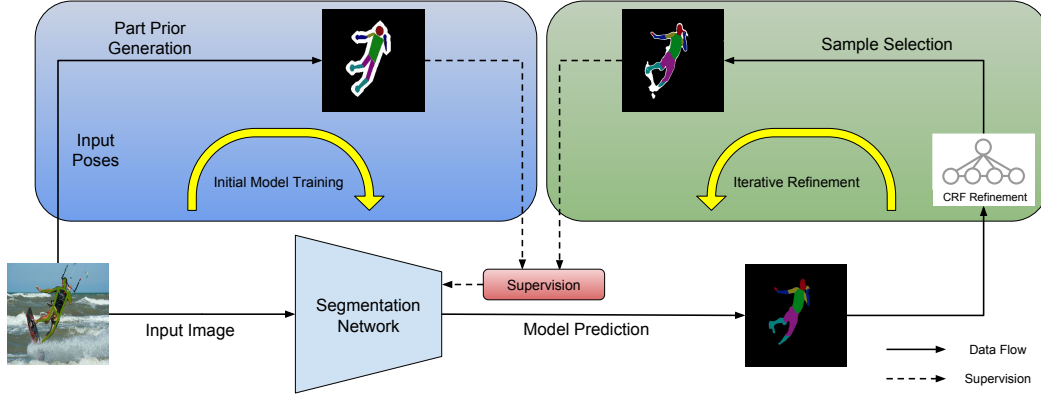


Fig. 2: The overall framework of the proposed method. The initial model training step (left blue box) aims to learn part structures, and is supervised by the part priors generated with input poses. The iterative refinement module (right green box) then refines the detailed shapes of predicted masks with image information. In segmentation masks, the colored, black and white regions denote the foreground body parts, backgrounds and uncertain regions respectively.

explored the conversion from sparse poses to the part prior masks. However, with poses only, it is challenging to directly generate part priors that are consistent with part boundaries reflected in images, since poses only contain body structure information with no cues about the detailed part shapes. Due to this inherent limitation, training the network directly with such part priors [3] generates unsatisfying masks with poor shape details. In semi-supervised studies, introducing a part prior refinement step [2] is effective, but extra manually labeled part masks are required.

In contrast to previous studies, we design part priors to have high precision on labeled regions, and we rely on the iterative refinement module to gradually recover the remaining uncertain boundary regions. As shown in Figure 1, we design part masks as ellipse or polygon templates. We use different templates for the upper body and lower body to better exploit the human structure prior. We show in section IV-D that such part priors well preserve the structure information, and do not require extra labels other than the poses.

Ellipse part priors. We draw ellipse $E(x, y, a, b, \alpha)$ on the estimated *head*, *upper arm* and *lower arm* regions based on the two corresponding poses P_i and P_j with the ellipse parameters:

$$\begin{aligned} x &= (P_j^x + P_i^x)/2 & y &= (P_j^y + P_i^y)/2 \\ a &= c_a * d(P_i, P_j) & b &= c_b * d(P_i, P_j) \\ \alpha &= \arctan((P_j^y - P_i^y) / (P_j^x - P_i^x)), \end{aligned} \quad (1)$$

where d is the $L2$ distance between the two keypoints and (x, y) is the center of the generated ellipse. c_a and c_b are pre-defined scale factors. Typical parameters for c_a and c_b are 0.6 and 0.22. Although the sizes and shapes of body parts vary, experiments in section IV-D show that the final results are robust to the hyper-parameter selection. The keypoints on the *shoulder* and *elbow*, the *elbow* and *wrist*, the *neck* and *head*, are paired to generate the ellipse part priors for the *upper arm*, *lower arm*, and *head*, respectively.

Polygon part priors. We draw polygons on the *torso*, *upper leg* and *lower leg* regions as part priors. The *torso* prior is generated by connecting the *left/right shoulder* and *left/right hips* to generate a quadrilateral. For the *leg* regions, the width of the leg is estimated to be half (*upper leg*) or one third (*lower leg*) of the distance from the *left/right hip* to the hip center. We then shift the keypoints of the *hip* and *knee* or the *knee* and *ankle* horizontally left and right to generate four points based on the estimated width, and generate the quadrilateral part priors for the *upper leg* or *lower leg*.

Estimated foregrounds. Furthermore, we generate estimated foreground masks based on part priors with the image dilation technique. The estimated foreground contains the generated part priors and uncertain blank regions. The size of the dilation matrix is proportional to the height of the poses. As shown in Figure 1, the generated part priors y contain three regions: estimated background B indicated by black, uncertain blank regions U indicated by white, and estimated foreground regions F indicated by other colors.

Missing keypoints. Although ground-truth poses are used for part prior generation, certain keypoints can still be missing because of viewpoints or being out of image boundary. We design recovery rules based on poses' spatial priors to alleviate this problem. For example, we estimate the location of the *neck* as the middle point of left and right *shoulder*. We draw the quadrilateral part priors for the *torso* as a parallelogram when one of four required keypoints is missing. Although the recovery step can not completely solve the inherent limitation of incomplete poses, it provides more part prior information and improves the final performance.

Overlapping and occlusion. The other inherent limitation of generating part priors with sparse keypoints is overlapping and occlusion. First, each individual's body parts overlap due to the viewpoint and pose structures. With only 2D keypoint information, it is impossible to perfectly recover the depth order. We make such an assumption that is correct in most

cases empirically: the lower arms/legs are in front of the upper arms/legs, the arms are in front of the legs, and the limbs are in front of the torso and head. Second, there also exist occlusions among different people. We rank the order with the number of visible keypoints, *i.e.*, we assume the person with the most annotated keypoints is in the front of others.

In summary, the scale, location, and orientation of generated parts are calculated based on corresponding keypoints. We do not try to generate part priors that fit perfectly with the image, as this cannot be accomplished. Instead, the objective is to learn a reliable initial prediction, which provides cues for part structures. The simple design also allows for an easy adaptation to other part segmentation tasks, as shown in section IV-E. We introduce the initial network training together with iterative refinement in section III-C.

C. Training Objective and Iterative Refinement

In this section, we first discuss the training objective in both modules. We then introduce the iterative refinement module.

Training objective. In each training iteration, the objective is to minimize the loss function L by learning parameters θ in the segmentation model $f(I; \theta)$. The weakly supervised loss function L consists of the structure loss L_s and the full mask loss L_m . The objective function is:

$$\min_{\theta} \sum_{I \in D} L_s(f(I; \theta)) + L_m(f(I; \theta)) * w_m. \quad (2)$$

In part structure learning stage with the pose generated part priors, we propose the structure loss L_s in the format of partial cross-entropy loss. L_s is only calculated on the confident foreground and background regions, *i.e.* $F \cup B$:

$$L_s = \sum_{i \in F \cup B} \sum_{c=0}^C \ell(f_i(c), y_i(c)), \quad (3)$$

where $f_i(c)$ is the model prediction score for class c at pixel i , y is the part prior generated from poses and ℓ is the per-pixel cross-entropy loss. With incomplete but highly confident priors on each body part, the network generates initial predictions with good part structures.

The full human mask also contains information related to part segmentation and is widely available. Furthermore, several datasets [14], [22] contain large-scale existing annotations for both poses and full masks. Because of this, a full mask loss L_m is proposed to utilize the extra weak supervision. With full mask annotation M , a binary cross-entropy mask loss is calculated on the whole image:

$$L_m = \sum_{i \in I} \sum_{fg=0}^1 \ell(\hat{M}_i(fg), M_i(fg)), \quad (4)$$

where foreground-background prediction \hat{M}_i is generated by the predicted background class probability $f_i(0)$.

Iterative refinement. To recover the blank uncertain regions and get better mask details, we propose the iterative refinement module. Although poses do not contain shape information,

we show that the mask shapes can be inferred from low-level pixel similarities. In the proposed iterative refinement module, we first refine the initial part prior supervised mask predictions on the training set with CRF. We adopt the dense CRF [17] in the iterative refinement step, where an appearance kernel and a smoothness kernel are used as the pairwise term. The refined masks show better prediction details by inferring pixel RGB similarities in the CRF. However, the downside of the refined masks is that the part structure information can be mistaken. For example, the predicted mask of *arm* might be incorrectly labeled as *torso* after the CRF refinement. Therefore, we propose to train the network jointly with initial part priors and generate new masks to include both structure and detail information in the iterative refinement.

Because of the variances in the complexity of the scene in image I and the quality of the corresponding predicted masks \hat{y}_i , different image label pairs contribute unequally towards segmentation model learning. Incorrectly predicted masks become noises in the next iteration and therefore deteriorate the model performance. Inspired by self-paced learning, we alleviate the problem with training sample selection during the iterative refinement. We follow a previous study [39] to discard unreliable predictions with a probability p_i in Eq. 5, and skip samples that have low prediction confidences in the next training iteration:

$$p_i = \max(0, 2 - \exp(\bar{f})) \quad (5)$$

where \bar{f} is the averaged pixel-wise prediction confidence over foreground regions:

$$\bar{f} = \frac{1}{N_F} \sum_{i \in F} \left(\max_{c=1 \dots C} f_i(c) \right) \quad (6)$$

IV. EXPERIMENTS

In this section, we first introduce the datasets used for experiments. We then compare the proposed body part segmentation methods to other state-of-the-art methods. Extensive discussions are conducted on various aspects of the method. Finally, we show the extension of the framework to other point supervised part segmentation tasks such as face parsing. The proposed framework is generally applicable for many segmentation methods, and we adopt Deeplab [18] as the base segmentation model in this study.

A. Datasets

Body part segmentation datasets. The *Pascal-Person-Part dataset* [8], [1] contains annotations for 14 human joints and 6 body parts. The annotated body parts are Head, Torso, Upper/Lower Arms, and Upper/Lower Legs. The total 3,533 images are split into 1,716 images for training and 1,817 images for testing. The *MSCOCO dataset* [14] contains over 150K human instance annotations with 1.7 million labeled keypoints. In this study, the pose annotations on 31K training images are used as the extra training data. We do not use the *LIP dataset* [22] in this study, because the extra clothing and object classes are unrelated to body part segmentation.

TABLE I: The body part segmentation performance on Pascal-Person-Part compared to the state-of-the-art. The best weakly supervised performances are highlighted with bold and the best fully supervised performances are marked with underlines. We show the weakly supervised performance with and without using extra data from MSCOCO. More discussions regarding the data size are in Table III.

Methods	Supervision	mIoU	Head	Torso	U-Arm	L-Arm	U-Leg	L-Leg	Bkg
LIP [22]	Fully	59.36	83.3	62.4	47.8	45.6	42.3	39.5	94.7
LG-LSTM [40]	Fully	57.97	82.7	61.0	45.4	47.8	42.3	38.0	88.6
Graph LSTM [41]	Fully	60.16	82.7	62.7	46.9	47.7	<u>45.7</u>	40.9	94.6
DeepLab [18]	Fully	<u>63.64</u>	<u>84.6</u>	<u>66.9</u>	<u>56.0</u>	<u>54.2</u>	<u>45.5</u>	<u>43.4</u>	<u>94.9</u>
Ours (Part prior)	Weakly	40.91	54.9	35.1	34.2	32.4	19.0	24.5	86.2
FastNet [3]	Weakly	42.11	61.6	37.8	32.7	29.2	20.8	25.0	87.1
Ours (Part prior supervision)	Weakly	43.91	50.6	47.0	31.8	29.5	29.3	27.9	87.0
Ours (Base, only PASCAL data)	Weakly	53.54	76.1	54.8	39.6	36.5	37.9	36.1	93.9
Ours (Iter, only PASCAL data)	Weakly	54.72	76.7	55.6	40.2	38.4	38.6	38.9	94.6
Ours (Base, with COCO data)	Weakly	60.35	78.3	59.6	46.6	45.4	46.6	50.0	95.8
Ours (Iter, with COCO data)	Weakly	62.05	79.6	62.0	48.1	48.5	48.7	51.8	95.8

Face parsing datasets. The *Helen dataset* [42] provides 194-point dense facial landmark annotations. A 10-class pixel-level annotation is defined and generated by Smith *et al.* [19]. The dataset contains 2,000 training images and 330 testing images. We follow the split in previous studies [19] and evaluate the face parsing task on the Helen testing set.

The *AFLW dataset* [20] contains around 25K annotated faces. Each face has 21 labeled facial landmarks. We filter out a 6K image subset that contains images with front-view faces. The AFLW dataset is used as the weakly supervised training data for the face parsing task evaluated on Helen.

B. Body Part Segmentation Results

Weakly supervised results. Table I reports the body part segmentation results on the Pascal-Person-Part dataset. The **top portion** of the table contains the numbers of several fully supervised state-of-the-art parsing methods [22], [40], [41], [18]. In order to separate the improvements brought by the proposed weakly supervised framework and that by advanced network structures, we use the same network [18] when comparing fully and weakly supervised methods. Therefore, we refer to the number generated by Deeplab [18] as the fully supervised results in the following discussions. The results of the compared weakly supervised baselines are shown in the **middle** of Table I. “Ours-part prior” directly evaluates the generated part priors as final predictions. “Ours-part prior supervision” takes the generated part priors as full supervisions and train the parsing network with the per-pixel cross-entropy loss [43]. Finally, the four rows at the **bottom** are the variations of our approach.

When iterative refinement and extra pose annotations from MSCOCO are used, the proposed weakly supervised approach “Ours-Iter-with COCO” achieves a mIoU of 62.05%, which is fairly comparable to the 63.64% acquired by the fully supervised methods [18] with the same network structure. We find there are two major reasons for the improvement. First, the proposed weakly supervised part segmentation framework makes learning from poses feasible, and improves the weakly supervised baseline performance by 10.8% in mIoU (cf. “Part prior supervision” and “Iter, only PASCAL data” in Table I). Second, the successful exploitation of extra weak supervisions further improves the performance by 7.3% (cf. “Iter, only PASCAL data” and “Iter, with COCO data”).

Semi-supervised results. The proposed framework also performs well under the semi-supervised setting. For semi-supervised learning, we replace part priors on Pascal-Person-Part with ground-truth part masks. All other settings remain the same as the weakly supervised experiment. Without learning specific adaptation networks to close the gap between different supervision types and domains, the proposed framework still generates a mIoU of 68.88% that outperforms the state-of-the-art (67.60% mIoU) [2].

C. Qualitative results analyses

In this section, we analyze the success and failure cases of our model as well as the compared methods to show the advantages and limitations of the proposed weakly supervised parsing framework. Qualitative results are shown in Figure 3. The first column is the input image with pose annotations visualized. The second to the forth columns show the ground-truth masks, weakly supervised baseline “Ours-part prior supervision”, our weakly supervised method “Ours-Base-with COCO” and “Ours-Iter-with COCO” respectively. The right two columns include the fully supervised results [18] and our semi-supervised results.

We show empirically from the qualitative results that the major advantage of using weak supervisions is to generate better part structures. For example in the second row, our weakly supervised method correctly generates masks for the *lower leg* regions, which are mispredicted as *background* by the fully supervised method. Similar examples can be observed in the second, third, and forth row of Figure 3. This observation is also applicable to semi-supervised results.

We then show the effectiveness of the iterative refinement. As shown in the fifth column, the refined results contain more accurate local details compared to the ones without iterative training in the fourth column. One clear example is the *neck* regions in the first row, which now have more precise boundaries.

Failure cases of our approach. Figure 4 shows the failure cases of our model. We observe three types of common failures: 1) The model might incorrectly predict occluded regions as one of the body parts instead of the background. For example, the dog in the first row of Figure 4. 2) Part priors can not be generated when corresponding joints are occluded or out of image boundaries. Because of this, such regions might

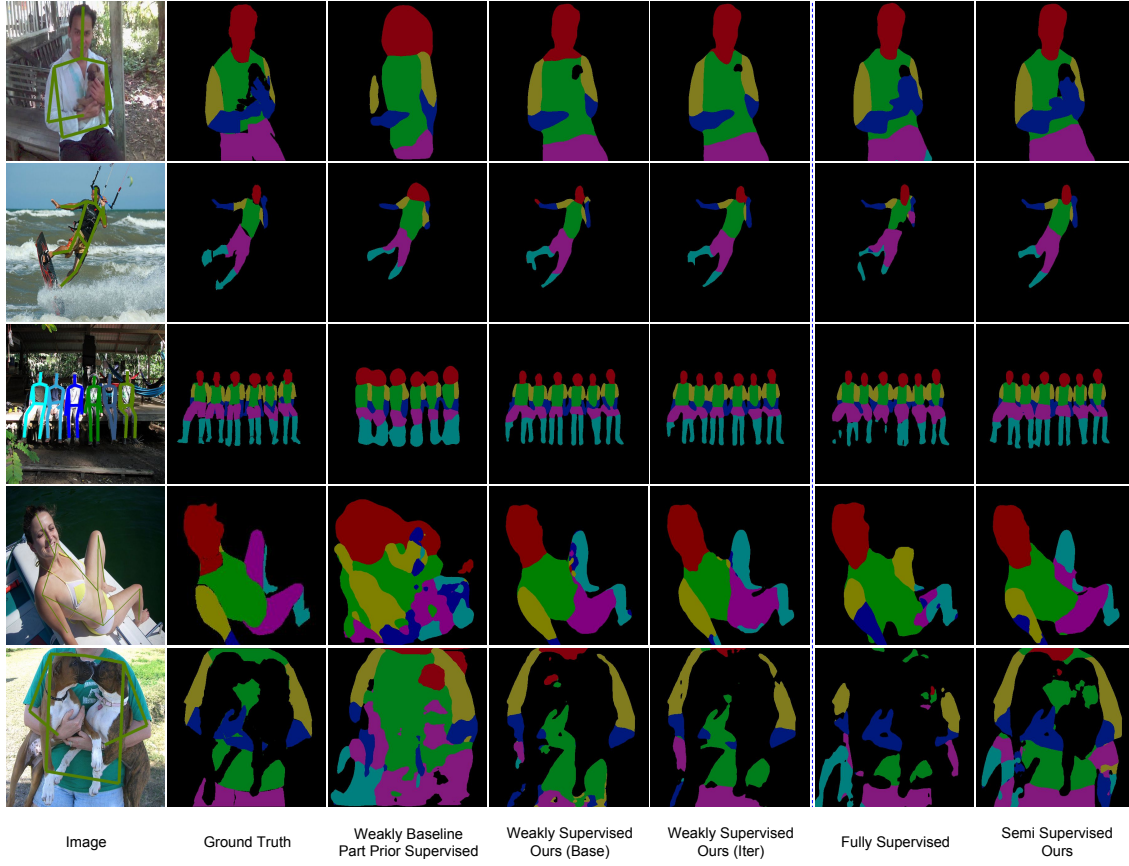


Fig. 3: The qualitative results on Pascal-Person-Part. The middle three columns are weakly supervised results. The right two columns utilize part segmentation annotations, i.e. fully and semi-supervised. Additional qualitative results can be found in the following link¹.

TABLE II: Comparison on different part prior generation methods. “Our part prior, w recovery” is our results to be compared to.

Methods	mIoU
Skeleton label map [1]	50.34
Our part prior, w/o recovery	52.77
Our part prior, w recovery	54.72
Our part prior, large param.	51.94
Our part prior, small param.	52.37
Our part prior, ideal param.	58.31

be mispredicted as background. For example, the *leg* regions in the second row. 3) Challenging cases with tiny body parts or complicated scenes might also fail the model, *e.g.*, the third row of Figure 4.

D. Discussions

Different part prior generation methods. We compare different part prior generation methods in Table II to show the superiority and robustness of the proposed part prior. In comparison, we report the performance of iterative training with only Pascal-person-part data. In the **top** portion of Table II, we compare our part prior generation methods to the skeleton label map [1]. The proposed part prior generation method (54.72% mIoU) outperforms the skeleton label map

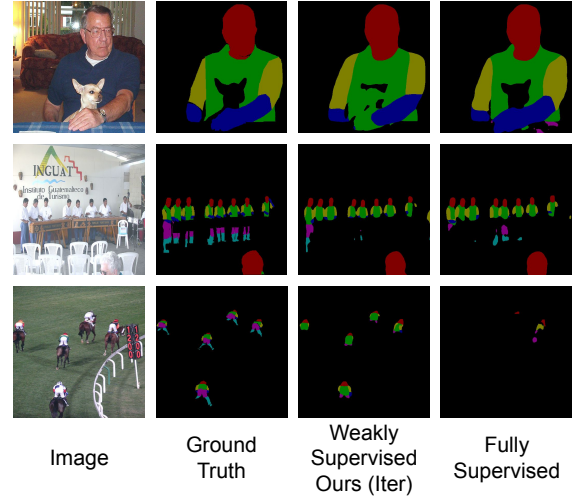


Fig. 4: The common failures of our weakly supervised method.

(50.34% mIoU) [1], which draws a stick between neighboring keypoints. In the **middle** of the table, we validate the effectiveness of overlapping recovery introduced in section III-B. The proposed recovery method improves mIoU by 2%.

With various body part sizes and shapes, it is impossible to select hyper-parameters that generate perfect part priors.

¹Additional qualitative results:
http://cs.rochester.edu/u/zyang39/weakly_parsing/pascal_visu.html

TABLE III: Weakly and semi supervised body part segmentation performance, trained with different amounts of extra weak annotations.

Methods	Pascal	COCO	mIoU
Ours (Weakly)	1.7K	0K	54.72
Ours (Weakly)	1.7K	10K	59.72
Ours (Weakly)	1.7K	31K	62.05
Fully	1.7K (part masks)	0K	63.64
Ours (Semi)	1.7K (part masks)	10K	67.68
Ours (Semi)	1.7K (part masks)	31K	68.88

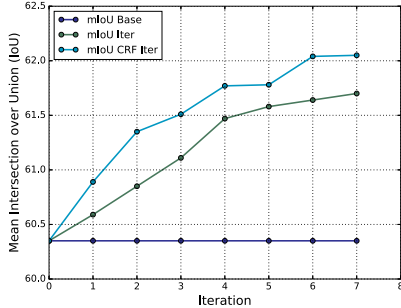


Fig. 5: The mIoU on Pascal-Person-Part during the iterative training.

At the **bottom** of the table, we show empirically that the framework is robust against different shape hyper-parameters. The upper bound of hyper-parameter selection is shown in “Our part prior, ideal parameters,” where ground-truth masks are used to fit the hyper-parameters. In the “large parameters” and “small parameters” experiment, we increase and decrease the shape hyper-parameters by 50%, respectively. By including the uncertain regions in part priors, the framework is robust against poor hyper-parameters, and the results remain better than other part prior generation methods [1].

The influence of training data size. As shown in Table III, we evaluate the influence of data size and supervision types. We train the proposed framework with different amounts of full annotations, weak annotations, or a mix of both. In fully- and semi-supervised experiments, we use the part mask annotations on Pascal-Person-Part that contain 1.7K images in total. In weakly supervised experiments, we use the poses and full masks from MSCOCO and Pascal-Person-Part. There are 31K images with poses and full masks on MSCOCO and 1.7K on Pascal-Person-Part. We evaluate all experiments on the testing set of Pascal-Person-Part. In Table III, we compare the weakly and semi supervised performances when using 1) no COCO data, 2) a subset of 10K COCO data, 3) and all available 31K COCO data. Our framework achieves significant improvements when extra weak supervision is adopted, *i.e.*, more than 5 mIoU for both the weakly and semi-supervised settings. The good performance is achieved by the successful exploitation of extra weak annotations, which have not been previously leveraged.

Iterative training. We show the effectiveness of iterative training in this section. As shown in Figure 5, iterative training brings steady improvements in mIoU during the first few iterations. As expected, observations show that the improvements are mainly achieved by generating better predictions on boundary pixels. Besides, conducting CRF refinement during

TABLE IV: The face parsing performance on the Helen dataset.

Methods	Supervision	mIoU
Baseline (Part Prior Supv.)	Weakly	42.54
Ours (Base)	Weakly	47.21
Ours (Iter)	Weakly	54.05
Fully Supervised	Fully	68.48
Ours (Semi)	Semi	73.95

iterative training brings an extra gain in performance.

Training without foreground masks. Besides using both poses and full masks, we show that the proposed framework also generates a good performance with only pose supervisions. Without full masks, the model achieves a mIoU of 53.62%, which is significantly better than the weakly supervised baselines (43.91% mIoU). Under the semi-supervised setting, the proposed framework achieves a mIoU of 66.68%.

E. Extensions to other Part Segmentation Tasks

Furthermore, the proposed pipeline can be easily adapted to other point-supervised part segmentation tasks, such as hand parsing, face parsing, and general object part segmentation. Such extensions could benefit other human-centered tasks such as emotion recognition [44]. As an example, we evaluate the framework on face parsing. In the weakly supervised face parsing task, we adopt the sparse 21 point facial landmarks as supervision. Similar to body part segmentation, we convert landmarks into part priors by drawing polygons. The facial landmark annotations from AFLW [20] are used to train the face parsing model. The face part definition follows previous studies [19] with two keypoints unrelated classes *lips* and *hair* merged. The method is evaluated on Helen [42], [19] that has dense face parts annotations.

As shown in Table IV, the proposed framework achieves a large improvement compared to the weakly supervised baseline. Similar to body part segmentation, the semi-supervised results with more data outperform the fully supervised baseline by a large margin. Figure 6 shows the qualitative results. The results show that the proposed framework can effectively utilize the abundant existing weak supervision to learn models when no full supervision is available, or combine with a small number of full annotations to further improve the performance.

V. CONCLUSION

To harvest the existing abundant weak annotations, we propose the first weakly supervised body part segmentation framework. The framework first uses pose-based part priors to learn the coarse locations of each part. The iterative refinement module then gradually recovers predictions on uncertain regions. The proposed framework works well in both weakly and semi-supervised settings by effectively using the extra weak annotations, which have not been previously leveraged. The weakly supervised performance with sufficient weak annotations is comparable to the fully supervised results with the same network, while the semi-supervised results outperform the state-of-the-art. Furthermore, we show that the framework can be applied to other weakly supervised part segmentation tasks with promising results on face parsing.

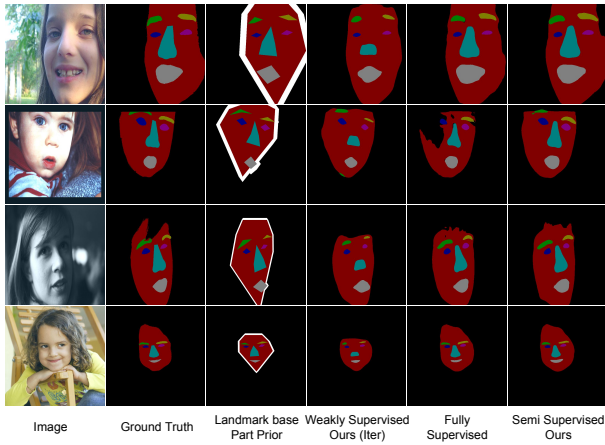


Fig. 6: The qualitative results on the Helen dataset.

ACKNOWLEDGMENT

This work is supported in part by NSF awards IIS-1704337, IIS-1722847, and IIS-1813709, Twitch Fellowship, as well as our corporate sponsors.

REFERENCES

- [1] F. Xia, P. Wang, X. Chen, and A. L. Yuille, "Joint multi-person pose estimation and semantic part segmentation," in *CVPR*, 2017.
- [2] H.-S. Fang, G. Lu, X. Fang, J. Xie, Y.-W. Tai, and C. Lu, "Weakly and semi supervised human body part parsing via pose-guided knowledge transfer," *CVPR*, 2018.
- [3] M. Zolfaghari, G. L. Oliveira, N. Sedaghat, and T. Brox, "Chained multi-stream networks exploiting pose, motion, and appearance for action classification and detection," in *ICCV*. IEEE, 2017, pp. 2923–2932.
- [4] M. M. Kalayeh, E. Basaran, M. Gökmen, M. E. Kamasak, and M. Shah, "Human semantic parsing for person re-identification," in *CVPR*, 2018, pp. 1062–1071.
- [5] G. Balakrishnan, A. Zhao, A. V. Dalca, F. Durand, and J. Guttag, "Synthesizing images of humans in unseen poses," in *CVPR*, June 2018.
- [6] K. Yamaguchi, M. Hadi Kiapour, and T. L. Berg, "Paper doll parsing: Retrieving similar styles to parse clothing items," in *ICCV*, 2013, pp. 3519–3526.
- [7] S. Liu, Y. Sun, D. Zhu, G. Ren, Y. Chen, J. Feng, and J. Han, "Cross-domain human parsing via adversarial feature and label adaptation," in *AAAI*, 2018.
- [8] 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, pp. 1971–1978.
- [9] 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*. Springer, 2016, pp. 648–663.
- [10] Y. Yang and D. Ramanan, "Articulated pose estimation with flexible mixtures-of-parts," in *CVPR 2011*. IEEE, 2011, pp. 1385–1392.
- [11] X. Nie, J. Feng, Y. Zuo, and S. Yan, "Human pose estimation with parsing induced learner," in *CVPR*, 2018, pp. 2100–2108.
- [12] Z. Yang, Y. Li, J. Yang, and J. Luo, "Action recognition with spatio-temporal visual attention on skeleton image sequences," *IEEE Transactions on Circuits and Systems for Video Technology*, 2018.
- [13] Z. Yang, Y. Li, J. Yang, and J. Luo, "Action recognition with visual attention on skeleton images," in *ICPR*, 2018.
- [14] 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*. Springer, 2014, pp. 740–755.
- [15] A. Bearman, O. Russakovsky, V. Ferrari, and L. Fei-Fei, "What's the point: Semantic segmentation with point supervision," in *ECCV*. Springer, 2016, pp. 549–565.
- [16] D. Lin, J. Dai, J. Jia, K. He, and J. Sun, "Scribblesup: Scribble-supervised convolutional networks for semantic segmentation," in *CVPR*, 2016, pp. 3159–3167.
- [17] P. Krähenbühl and V. Koltun, "Efficient inference in fully connected crfs with gaussian edge potentials," in *Advances in neural information processing systems*, 2011, pp. 109–117.
- [18] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," in *ECCV*, 2018.
- [19] B. M. Smith, L. Zhang, J. Brandt, Z. Lin, and J. Yang, "Exemplar-based face parsing," in *CVPR*, 2013, pp. 3484–3491.
- [20] K. Martin, W. Paul, M. R. Peter, and B. Horst, "Annotated Facial Landmarks in the Wild: A Large-scale, Real-world Database for Facial Landmark Localization," in *Proc. First IEEE International Workshop on Benchmarking Facial Image Analysis Technologies*, 2011.
- [21] K. Gong, X. Liang, Y. Li, Y. Chen, M. Yang, and L. Lin, "Instance-level human parsing via part grouping network," in *ECCV*, September 2018.
- [22] K. Gong, X. Liang, D. Zhang, X. Shen, and L. Lin, "Look into person: Self-supervised structure-sensitive learning and a new benchmark for human parsing," in *CVPR*, 2017.
- [23] Y. Luo, Z. Zheng, L. Zheng, G. Tao, Y. Junqing, and Y. Yang, "Macro-micro adversarial network for human parsing," in *ECCV*, 2018, pp. 424–440.
- [24] K. Gong, X. Liang, Y. Li, Y. Chen, M. Yang, and L. Lin, "Instance-level human parsing via part grouping network," in *ECCV*, 2018, pp. 770–785.
- [25] X. Nie, J. Feng, and S. Yan, "Mutual learning to adapt for joint human parsing and pose estimation," in *ECCV*, 2018.
- [26] K. Gong, Y. Gao, X. Liang, X. Shen, M. Wang, and L. Lin, "Graphonomy: Universal human parsing via graph transfer learning," in *CVPR*, 2019, pp. 7450–7459.
- [27] W. Wang, Z. Zhang, S. Qi, J. Shen, Y. Pang, and L. Shao, "Learning compositional neural information fusion for human parsing," in *Proceedings of the IEEE International Conference on Computer Vision*, 2019, pp. 5703–5713.
- [28] M. Tang, F. Perazzi, A. Djelouah, I. Ben Ayed, C. Schroers, and Y. Boykov, "On regularized losses for weakly-supervised cnn segmentation," in *ECCV*, September 2018.
- [29] R. Hu, P. Dollár, K. He, T. Darrell, and R. Girshick, "Learning to segment every thing," in *CVPR*, 2018.
- [30] G. Papandreou, L.-C. Chen, K. P. Murphy, and A. L. Yuille, "Weakly- and semi-supervised learning of a deep convolutional network for semantic image segmentation," in *ICCV*, December 2015.
- [31] T. Remez, J. Huang, and M. Brown, "Learning to segment via cut-and-paste," in *ECCV*, September 2018.
- [32] S. Kwak, S. Hong, B. Han *et al.*, "Weakly supervised semantic segmentation using superpixel pooling network," in *AAAI*, 2017, pp. 4111–4117.
- [33] Y. Wei, X. Liang, Y. Chen, X. Shen, M.-M. Cheng, J. Feng, Y. Zhao, and S. Yan, "Stc: A simple to complex framework for weakly-supervised semantic segmentation," *TPAMI*, vol. 39, no. 11, pp. 2314–2320, 2017.
- [34] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, "Learning deep features for discriminative localization," in *CVPR*, 2016.
- [35] A. Kolesnikov and C. H. Lampert, "Seed, expand and constrain: Three principles for weakly-supervised image segmentation," in *ECCV*. Springer, 2016, pp. 695–711.
- [36] Q. Li, A. Arnab, and P. H. Torr, "Weakly- and semi-supervised panoptic segmentation," in *ECCV*, September 2018.
- [37] G. Seguin, K. Alahari, J. Sivic, and I. Laptev, "Pose estimation and segmentation of multiple people in stereoscopic movies," *IEEE T-PAMI*, vol. 37, no. 8, pp. 1643–1655, 2014.
- [38] S. Zheng, S. Jayasumana, B. Romera-Paredes, V. Vineet, Z. Su, D. Du, C. Huang, and P. H. Torr, "Conditional random fields as recurrent neural networks," in *CVPR*, 2015, pp. 1529–1537.
- [39] Q. You, J. Luo, H. Jin, and J. Yang, "Robust image sentiment analysis using progressively trained and domain transferred deep networks," in *AAAI*, 2015, pp. 381–388.
- [40] 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, pp. 3185–3193.
- [41] X. Liang, X. Shen, J. Feng, L. Lin, and S. Yan, "Semantic object parsing with graph lstm," in *ECCV*. Springer, 2016, pp. 125–143.
- [42] V. Le, J. Brandt, Z. Lin, L. Bourdev, and T. S. Huang, "Interactive facial feature localization," in *ECCV*. Springer, 2012, pp. 679–692.
- [43] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *CVPR*, 2015, pp. 3431–3440.
- [44] Z. Yang, Y. Zhang, and J. Luo, "Human-centered emotion recognition in animated gifs," in *ICME*. IEEE, 2019.