Weakly Supervised Body Part Parsing with Pose based Part Priors
Zhengyuan Yang,Yuncheng Li, Linjie Yang, Ning Zhang, and Jiebo Luo

Abstract—Human body part parsing refers to the task of predicting the semantic segmentation mask for each body part. Fully supervised body part parsing methods achieve good performances, but require an enormous amount of effort to annotate part masks for training. In contrast to high annotation costs required 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 parsing framework. The basic idea is to train a parsing network with pose generated part priors that has blank uncertain regions on estimated boundaries, and use an iterative refinement module to generate new supervision and predictions on these regions. When sufficient extra weak supervisions are available, our weakly-supervised results (62.6% mIoU) on Pascal-Person-Part are comparable to the fully supervised state-of-the-art results (63.6% mIoU). Furthermore, in the extended semi-supervised setting, the proposed framework outperforms the state-of-art methods. In addition, we show that the proposed framework can be extended to other keypoint-supervised part parsing tasks such as face parsing.

Index Terms—Part parsing, weakly supervised learning, segmentation.

I. INTRODUCTION

Body part parsing [1], [2] is a fundamental task for understanding human activity in visual content. The objective of body part parsing is to generate pixel-level labels for semantic body parts. It is related to human parsing [3], but focuses more on predicting part masks that directly reflects body structures. Body part parsing helps various vision tasks such as action recognition [4], person re-id [5], and image generation [6]. It has also been applied to various application domains such as online shopping [7] and surveillance systems [8]. Previous studies on part parsing [9], [10], [1], [2] are conducted mostly in a fully supervised setting and thus require manually annotated part masks as supervision during training. However, the number of annotations for body part parsing is often limited, due to high annotation costs of labeling the pixel-level part masks. The Pascal-Person-Part dataset [9], [1] is the largest body part parsing dataset by far but contains only 3.5K images in total.

On the other hand, huge amounts of available weak annotations such as human poses and full body masks already exist, and contain information related to the part parsing task. In particular, human poses contain cues for high-level human structures and the coarse spatial locations of human parts [11], [12], [13]. 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 parsing. 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 parsing 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 human poses and full masks as an alternative supervision for training the body part parsing task. The initial step is to generate part priors with poses, which include blank uncertain regions on the estimated part boundaries. The part prior generation convert poses into initial part masks by exploiting the prior knowledge on human structures. The boundary regions are then gradually recovered by an iterative refinement module. The predictions are refined by the Conditional Random Field (CRF) [17] and are used as supervisions to iteratively improve the segmentation network.

To the best of our knowledge, this is the first feasible solution on weakly supervised body part parsing. 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 methods [2]. We show that the widely available weak supervisions improve the performance of body part parsing by a large margin. Furthermore, the part parsing framework can be extended to objects other than human body, where keypoint annotations are available. To validate the effectiveness of our approach, we also evaluate the weakly supervised framework on a face parsing task on the Helen dataset [19] and the AFLW dataset [20], and show promising results.

How does the proposed framework obtain the significant improvement and why can weak labels help the part parsing task? We show empirically in discussion (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. Results also show that the gain is mainly
obtained by reducing errors in body structure prediction, e.g. confusing arms with legs. Further, extensively experiments show the superiority and robustness of the proposed pose based part prior. The new problem formulation, the appreciable performance improvements, and the effective mechanisms to achieve the goal, constitute the main contributions of this work.

II. RELATED WORKS

Human body part parsing. Human body part parsing is the task of generating body part masks based on human structures. An Auto-Zoom Network [10] is proposed to focus on certain body part regions. Xia et al. [1] propose to jointly conduct body part parsing 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 parsing based on pose similarities. Part parsing is related to human parsing [21], [23], [24], [25], as both tasks aim to predict pixel level semantic masks for human parts. The major difference is that body part parsing focuses on predicting part masks that directly reflects body structures, for example regions such as upper/lower arms. On the contrary, human parsing contains clothing and object classes such as sunglasses, hats, coats and etc. These classes are not directly related to human structures and are more appearance based.

Weakly supervised semantic segmentation. Our study is also related to weakly supervised semantic segmentation. Frequently used weak supervisions for semantic segmentation include scribbles [16], [26], bounding boxes [27], [28], [29], points [15] and image labels [30], [31], [32], [33]. Despite the promising results achieved on the semantic segmentation task, existing weakly supervised segmentation methods require the saliency of target regions in color space. Without stronger supervisions or revised methods, directly applying these previous studies onto more complicated scenarios, such as scene parsing [34] and part parsing, generates limited performance. To solve the body part parsing task in this study, we utilize the human structure knowledge with the proposed pose based part prior. Our study is also related to previous explorations on transforming body poses [11], [35].

III. METHODOLOGY

A. Overview

We first introduce the problem formulation of body part parsing. 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 parsing 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 \( L_n \in \{0, 1, \ldots, C\}^{H \times W} \) are manually labeled and the network is trained with the per-pixel cross entropy loss \( \ell \). 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 parsing 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 parsing framework without any pixel-level part annotations.

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 to only cover the most confident regions, i.e. to generate part priors that have high precision. The objective is to provide the cues of the coarse location of each part, instead of learning the detailed part shapes. 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 initial 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 initial part priors and the refined mask predictions are used as supervisions in the iterative training. Self-paced learning is further adopted to select reliable enough predictions as the supervision during iterative refinement. Furthermore, we find that end-to-end CRF refinement such as CRFs3D [36] 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], [4], [2] have explored the conversion from sparse poses to 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 [4] 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, part masks are designed as ellipse or polygon templates. Different templates are used 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:

\[
\begin{align*}
  x &= \frac{(P^x_j + P^x_i)}{2} \\
  y &= \frac{(P^y_j + P^y_i)}{2} \\
  a &= c_a \ast d(P_i, P_j) \\
  b &= c_b \ast d(P_i, P_j) \\
  \alpha &= \text{arctan}\left(\frac{(P^y_j - P^y_i)}{(P^x_j - P^x_i)}\right) 
\end{align*}
\]

(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 with respect 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.** Polygons are drawn 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 dilation matrix is proportional to the height of 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, and we complete 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 improve 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 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 exists occlusions among different people. We rank the order with number of visible keypoints, i.e. we assume the person with the most annotated keypoints comes in 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 provide cues for part structures. The simple design also allows for an easy adaptation to other part parsing 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 $\theta$ 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:

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

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. $F \cup B$:

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

where $f_i(c)$ is the model prediction score for class $c$ at pixel $i$, $y$ is the part prior generated from poses and $\ell$ 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 parsing and is widely available. Furthermore, a number of datasets [14], [3] 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:

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

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 $y_i$, different image label pairs contribute unequally towards segmentation model learning. Incorrect 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 [37] 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:

$$p_i = \max(0, 2 - \exp(\tilde{T})) \tag{5}$$

where $\tilde{T}$ is the averaged pixel-wise prediction confidence over foreground regions:

$$\tilde{T} = \frac{1}{N_F} \sum_{i \in F} \left( \max_{c=1...C} f_i(c) \right) \tag{6}$$

IV. Experiments

In this section, we first introduce the datasets used for experiments. We then compare the proposed body part parsing 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 parsing tasks such as face parsing. The proposed framework is general applicable for many segmentation methods, and we adopt Deeplab [18] as the base segmentation model in this study.

A. Datasets

Body part parsing datasets. The Pascal-Person-Part dataset [9], [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 [3] in this study, because the extra clothing and object classes are unrelated to body part parsing.

Face parsing datasets. The Helen dataset [40] 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 Parsing Results

Weakly supervised results. Table I reports the body part parsing 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 [3], [38], [39], [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 following discussions. The results of the compared weakly supervised baselines are shown in the middle of Table I. “Ours-part priot” directly evaluate 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 [41]. Finally, the four rows at the bottom are the variations of our own approach.

When iterative refinement and extra pose annotations from MSCOCO are used, the proposed weakly supervised approach “Ours-Iter-with COCO” achieves an 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 parsing framework makes learning from poses feasible, and improves the weakly supervised baseline performance by 10.8% in mIoU. Second, the successful exploitation of extra weak supervisions further improve the performance by 7.3%.

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 an 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 utilizing weak supervisions is the better predicted part structures. For example in the second row, our weakly supervised method correctly generates masks for the lower leg regions, which are predicted incorrectly 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 on 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 accurate boundaries.

Failure cases of our approach. Figure 4 shows 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 be incorrectly predicted 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 the comparison, we report the performance of iterative training
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 link.

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

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 |

keypoints. In the middle of the table, we validate the effectiveness of overlapping recovery introduced in section III-B. The proposed recovery method improves the mIoU by 2%.

With various body part sizes and shapes, it is impossible to select hyper-parameters that generate perfect part priors. In 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. The proposed framework is trained with different amounts
TABLE III: Weakly and semi supervised body part parsing 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.

of full annotations, weak annotations, or a mix of both. In full- 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. All experiments are evaluated 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), and 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 prediction on boundary pixels. Besides, conducting CRF refinement during 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 generates a good performance with only pose supervisions. Without full masks, the model achieves an 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 an mIoU of 66.68%.

E. Extensions to other Part Parsing Tasks

Furthermore, the proposed pipeline can be easily applied to other point-supervised part parsing tasks, such as hand parsing, face parsing, and general object part parsing. 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 parsing, 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 the lips and hair merged. The method is evaluated on Helen [40], [19] that has dense face parts annotations.

As shown in Table IV, the proposed framework achieves a significant improvement compared to the weakly supervised baseline. Similar to body part parsing, the semi-supervised results with more data outperform the fully supervised baseline by a large margin. The qualitative results are shown in Figure 6. The promising 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 amount 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 parsing 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 parsing tasks with promising results on face parsing.
ACKNOWLEDGMENT

This work is partially supported by NSF awards #1704337, #1722847, and #1813709.

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] 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.

[4] 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.

[5] 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.

[6] G. Balakrishnan, A. Zhao, A. V. Dalca, F. Durand, and J. Guttag, “Synthesizing images of humans in unseen poses,” in CVPR, June 2018.

[7] K. Yamaguchi, M. Hadi Kiapour, and T. L. Berg, “Paper doll parsing: Rendring similar styles to parse clothing items,” in ICCV, 2013, pp. 3519–3526.

[8] 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.

[9] 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.

[10] 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.

[11] Y. Yang and D. Ramanan, “Articulated pose estimation with flexible mixtures-of-parts,” in CVPR 2011. IEEE, 2011, pp. 1385–1392.

[12] X. Nie, J. Feng, Y. Zuo, and S. Yan, “Human pose estimation with parsing induced learner,” in CVPR, 2018, pp. 2100–2108.

[13] 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.

[14] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollar, and C. L. Zitnick, “Microsoft coco: Common objects in context,” in ECCV. Springer, 2014, pp. 740–755.

[15] A. Beamam, O. Russakovsky, V. Ferrari, and L. Fei-Fei, “Whats 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 NIPS. Curran, 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] Y. Luo, Z. Zheng, L. Zheng, G. Tao, Y. Junsung, and Y. Yang, “Macro-micro adversarial network for human parsing,” in ECCV, 2018, pp. 424–440.

[23] K. Gong, X. Liang, Y. Li, Y. Chen, M. Yang, and L. Lin, “Instance-level human parsing via part grouping network,” in Proceedings of the European Conference on Computer Vision (ECCV), 2018, pp. 770–785.

[24] X. Nie, J. Feng, and S. Yan, “Mutual learning to adapt for joint human parsing and pose estimation,” in ECCV, 2018.

[25] K. Gong, Y. Gao, X. Liang, X. Shen, M. Wang, and L. Lin, “Graphonomy: Universal human parsing via graph transfer learning,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 7450–7459.

[26] 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.

[27] R. Hu, P. Dollr, K. He, T. Darrell, and R. Girshick, “Learning to segment every thing,” in CVPR, 2018.

[28] 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.

[29] T. Remez, J. Huang, and M. Brown, “Learning to segment via cut-and-paste,” in ECCV, September 2018.

[30] S. Kwak, S. Hong, B. Han et al., “Weakly supervised semantic segmentation using superpixel pooling network,” in AAAI, 2017, pp. 4111–4117.

[31] 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.

[32] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, “Learning deep features for discriminative localization,” in CVPR, 2016, pp. 2921–2929.

[33] A. Kolesnikov and C. H. Lampert, “Seed, expand and constrain: Three principles for weakly-supervised image segmentation,” in ECCV. Springer, 2016, pp. 695–711.

[34] Q. Li, A. Arnab, and P. H. Torr, “Weakly- and semi-supervised panoptic segmentation,” in ECCV, September 2018.

[35] G. Seguin, K. Alahari, J. Sivic, and I. Laptev, “Pose estimation and segmentation of multiple people in stereoscopic movies,” IEEE transactions on pattern analysis and machine intelligence, vol. 37, no. 8, pp. 1643–1655, 2014.

[36] 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.

[37] 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.

[38] 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.

[39] Y. Li, X. Shen, J. Feng, L. Lin, and S. Yan, “Semantic object parsing with graph lstm,” in ECCV. Springer, 2016, pp. 125–143.

[40] V. Le, J. Brandt, Z. Lin, L. Bourdev, and T. S. Huang, “Interactive facial feature localization,” in ECCV. Springer, 2012, pp. 679–692.

[41] J. Long, E. Shelhamer, and T. Darrell, “Fully convolutional networks for semantic segmentation,” in CVPR, 2015, pp. 3431–3440.