

Keypoint Based Weakly Supervised Human Parsing

Zhonghua Wu, Guosheng Lin, Jianfei Cai

Nanyang Technological University, Singapore
{zhonghuawu, gslin, asjfcai}@ntu.edu.sg

Abstract

Fully convolutional networks (FCN) have achieved great success in human parsing in recent years. In conventional human parsing tasks, pixel-level labeling is required for guiding the training, which usually involves enormous human labeling efforts. To ease the labeling efforts, we propose a novel weakly supervised human parsing method which only requires simple object keypoint annotations for learning. We develop an iterative learning method to generate pseudo part segmentation masks from keypoint labels. With these pseudo masks, we train a FCN network to output pixel-level human parsing predictions. Furthermore, we develop a correlation network to perform joint prediction of part and object segmentation masks and improve the segmentation performance. The experiment results show that our weakly supervised method is able to achieve very competitive human parsing results. Despite our method only uses simple keypoint annotations for learning, we are able to achieve comparable performance with fully supervised methods which use the expensive pixel-level annotations.

Introduction

Semantic image segmentation is a fundamental task for image understanding. Human parsing, also known as human part segmentation, can be considered as a part-level image segmentation task. Human parsing aims to segment one person into different parts, which is a pixel labeling task and plays an important role in human analysis. Part segmentation or human parsing has recently attracted increasing attention in the research community (Lin et al. 2017; Chen et al. 2014; Liang et al. 2015; Wang et al. 2015; Xia et al. 2016). Human parsing stimulates various high-level vision understanding applications such as action recognition, human behavior analysis and video surveillance.

Conventional part segmentation methods require pixel-level annotations for training which usually involve excessive human labeling efforts.

To avoid this huge burden of pixel-level annotations, in this research we propose to use simple object keypoint annotation as supervision for learning human parsing models. Compared to pixel-wise part annotations, object keypoint

annotations are much easier to obtain which significantly reduces human labeling efforts. Fig. 1 illustrates the idea of our keypoint based weakly supervised human parsing framework and how it is different from conventional fully supervised methods.

Object keypoint annotations can be obtained from human labelling, e.g., keypoint annotations in human pose datasets (Xia et al. 2017; Lin et al. 2014), or from pre-trained human pose estimation models, e.g, Mask RCNN (He et al. 2017) and AlphaPose (Fang et al. 2017; Xiu et al. 2018). We demonstrate our proposed method is able to incorporate with human labeled keypoint annotations as well as the less accurate keypoint predictions generated by pre-trained human keypoint detection models to achieve pixel-level human part and object segmentation. In addition, considering there is a strong correlation between the whole object segmentation and the part segmentation. We develop a joint learning method to model such correlations and simultaneously output object and part segmentation masks. Particularly, we introduce a correlation block to model interaction between the part prediction and the object prediction, and it helps to improve the final part segmentation result.

Our main contributions are summarized as follows:

- We propose a weakly supervised method to ease the human labeling efforts for human parsing. We are able to achieve good pixel-level human part and object segmentation results using only simple object keypoint annotations as supervision for learning. The object keypoints can be obtained from human manual labeling or pre-trained object keypoint detectors. Our method significantly reduces human labeling efforts and achieves very competitive performance for human parsing.
- We propose an iterative learning method to generate accurate pseudo masks for parts and objects from object keypoint annotations. With such high-quality pseudo masks, we train a segmentation network to jointly predict pixel-level part and object segmentation masks.
- Due to the strong correlation between parts and objects, joint prediction is expected to benefit the segmentation performance. We propose a correlation network to simultaneously output part and object segmentation masks, and achieve improved results for part segmentation.

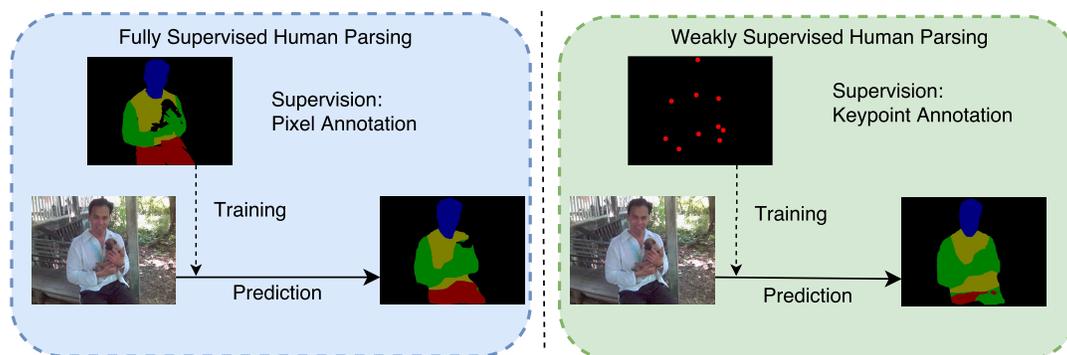


Figure 1: Comparison between fully supervised methods and our weakly supervised method. The left box (blue) describes traditional fully supervised human parsing, which requires expensive pixel-level annotations for training. The right box (green) illustrates our weakly supervised human parsing method which only requires simple object keypoint annotations for training. We are able to achieve comparable performance with fully supervised methods.

Related work

Our method is related to the research themes including weakly supervised segmentation, human part segmentation and pose estimation.

Weakly supervised segmentation

In the recent years, the development of deep convolutional neural networks (CNN) with advanced network structures such as VGG (Simonyan and Zisserman 2014a) and ResNet (He et al. 2016) have been widely used in many areas such as object detection and segmentation. The work in (Long, Shelhamer, and Darrell 2015) proposes fully convolutional neural networks (FCN) based on VGG network for semantic segmentation with end-to-end learning. The approach in (Chen et al. 2016a) introduces Atrous/dilated convolution and employs fully connected CRFs to improve the FCN method.

Conventional fully supervised segmentation requires pixel-wise mask annotations for training and it requires enormous human labeling effort which is usually excessively expensive. To ease the labeling efforts, a number of weakly supervised methods (Pathak et al. 2014; Vernaza and Chandraker 2017; Lin et al. 2016; Bearman et al. 2016; Dai, He, and Sun 2015; Papandreou et al. 2015; Shen et al. 2017; Zhang et al. 2018) have been proposed to employ weak supervision for learning segmentation models. For example, the methods in (Pathak et al. 2014; Papandreou et al. 2015; Shen et al. 2017) use image level labels for learning segmentation models; the work in (Bearman et al. 2016) use image level and point level information for model training. The work in (Dai, He, and Sun 2015) and (Lin et al. 2016) use box annotation and scribble annotation, respectively, as supervision for learning segmentation models.

Different from these existing studies, we focus on human part segmentation. We propose to use object keypoint annotation as supervision which is more challenging than using scribble or box-level supervision for learning segmentation models.

Human part segmentation

The work in (Chen et al. 2014) extends object segmentation to object part-level segmentation. It releases a PASCAL PART dataset which contains pixel-level part annotations. The work in (Wang et al. 2015) first attempts part segmentation on animals.

It uses fully-connected CRFs as post-processing to enhance the consistency between parts and objects.

The approach in (Chen et al. 2016b) proposes an attention model to fuse multi-scale prediction for part segmentation. The work in (Xia et al. 2016) uses the “auto-zoom” to build a hierarchical model to adapt the scales for objects and parts.

The method in (Li et al. 2017) extends single human parsing to multiple human parsing and demonstrates in real-world applications. The method in (Tsogkas et al. 2015) employs high-level information to improve part segmentation. The approach in (Lin et al. 2017) proposes a multi-path refinement network to achieve high resolution and accurate part segmentation.

Different from these fully supervised works, we focus on weakly supervised human parsing which only use human keypoint annotations rather than pixel level annotations.

Pose estimation

The work in (Cao et al. 2017) proposes Part Affinity Fields for human pose estimation. Then multi-stage pose estimation methods (Fang et al. 2017; He et al. 2017) use object detection and segmentation information to guide pose estimation prediction. The method in (Xia et al. 2017) jointly performs the multi-person pose estimation and the semantic part segmentation in a single image to enhance the performance of both segmentation and pose estimation.

In this work, we use human pose estimation method to generate pseudo human keypoint to refine or generate our human part segmentation.

Approach

Fig. 2 gives an overview of the proposed method. It consists of three parts: the human keypoint annotation part in

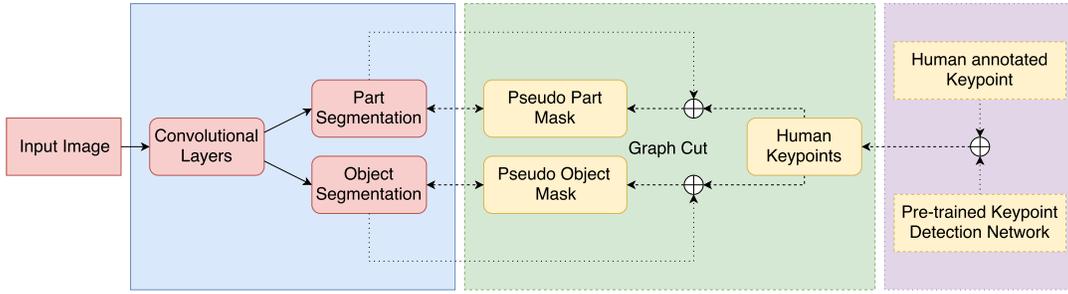


Figure 2: An overview of our weakly supervised method. The right part (purple) shows the object keypoint annotations which can be obtained from human labeling or pre-trained keypoint detectors such as MaskRCNN or AlphaPose. The middle part (green) illustrates our pseudo mask generation for parts and objects from keypoint annotations. The left part (blue) describes our FCN based segmentation network with two output branches for a joint and part learning. This segmentation network uses pseudo masks described in the middle part for training. In the test stage, only the trained segmentation network (left part) is applied for the part segmentation prediction.

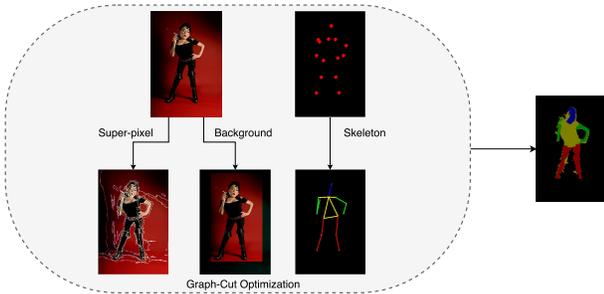


Figure 3: Illustration of the pseudo mask generation from keypoint annotations. Firstly, we generate super-pixels (Uijlings et al. 2013) of the input image (first bottom column). Secondly, we estimate background regions based on the location of keypoints (second bottom column). We treat the pixels which are 50 pixels far away from the human keypoints as the background. Thirdly, we connect the object keypoints to generate the skeleton (third bottom column). Finally, we construct a graph-cut model to generate the pseudo masks of parts.

the right, the pseudo mask generation in the middle, and the FCN based segmentation network in the left. In particular, we generate pseudo object and part masks from object keypoint annotations. Then the resulting pseudo masks are employed for training our FCN based joint segmentation networks for part and object mask prediction. In the following, we focus on elaborating our three specially designed processes, i.e. generating pseudo marks from keypoints, iteratively refining pseudo marks and correlation network for joint prediction.

Generating pseudo masks from keypoints

At the first step, we build a graphical model to generate pseudo masks of objects and parts from keypoint annotations. We generate super-pixels for our training images and construct a graph over the super-pixels. This problem can be formulated as an energy minimization problem. The energy

function is written as:

$$E = \sum_i \varphi_i(y_i) + \sum_{i,j} \varphi_{ij}(y_i, y_j), \quad (1)$$

where $\varphi_i(y_i)$ is the unary term indicating the labeling confidence for one super-pixel, and $\varphi_{ij}(y_i, y_j)$ is the pairwise term indicating the pairwise labeling confidence for a pair of neighboring super-pixels. Here $y \in \{1 \dots K\}$ denotes the part label which takes a value from one of the K labels.

We construct the unary term based on object keypoint annotations. As shown in Fig. 3, we connect object keypoints to generate a skeleton, and all pieces of the skeleton are assigned object part labels based on the types of object keypoints. A super-pixel overlapped with the skeleton will be assigned a part label, denoted by L , according to the overlapped skeleton pieces. If there are two skeletons across one super-pixel, the superpixel will be given the label of the skeleton with more overlapped pixels. We consider all part regions as confident foreground regions. We formulate the unary term cost function as:

$$\varphi_i(y_i) = \begin{cases} -\log(\frac{1}{|K|}) & \text{if } X_i \cap S = \emptyset; \\ 0 & \text{if } X_i \cap S \neq \emptyset, y_i = L_i; \\ a \text{ large value} & \text{if } X_i \cap S \neq \emptyset, y_i \neq L_i, \end{cases} \quad (2)$$

where X_i indicates a super-pixel and S indicates the skeleton. Here $X_i \cap S = \emptyset$ indicates a super-pixel X_i does not overlap with any pieces of the skeleton S , and likewise, $X_i \cap S \neq \emptyset$ indicates a super-pixel X_i overlaps with some pieces of the skeleton S . In the first case, i.e. the super-pixel does not overlap with the skeleton, the costs of all part categories are set to the same, with K indicating the total number of part categories.

Recall we assign a part label, denoted by L_i , to the super-pixel i based on the overlapped piece of the skeleton. In the cases when the super-pixel overlaps with the skeleton, if y_i equals to the assigned label L_i , we set the cost to 0; otherwise, we set a high cost value, e.g., 10^7 in our implementation.

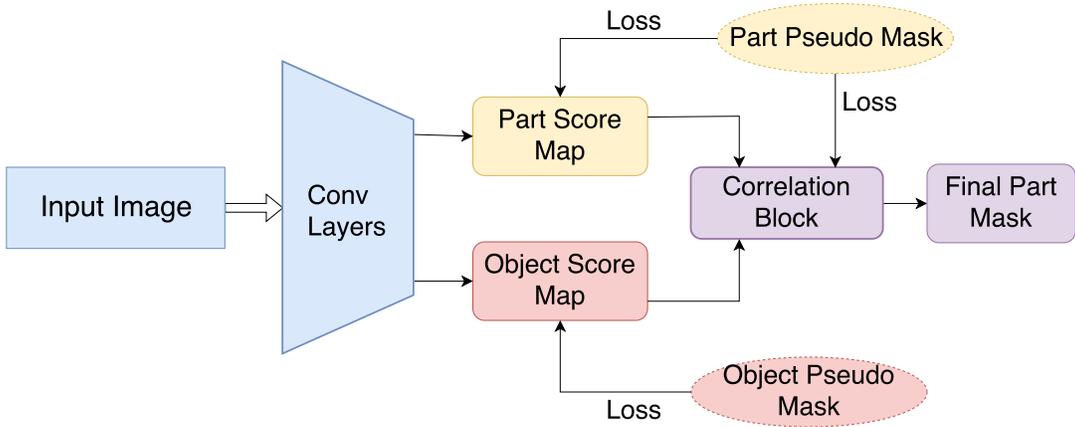


Figure 4: Illustration of our correlation network for joint learning of parts and objects. Our network contains two branches for part and object prediction. We introduce a correlation block to model the interaction between parts and objects, and thus to improve the final part segmentation.

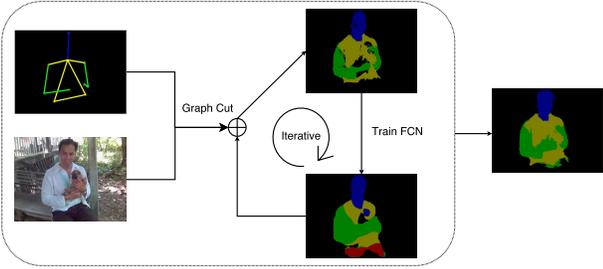


Figure 5: Illustration of our iteration refinement process for pseudo mask generation. We perform graph cut prediction and FCN training iteratively to improve pseudo masks. In the first iteration, we build a graph-cut model based on object keypoints to generate pseudo masks, and then we train FCN based segmentation network using the pseudo masks. In the next iteration, we jointly consider the keypoint annotations and the segmentation score map, generated by the trained FCN based model from the last iteration, to construct a new graph-cut model to construct better pseudo masks. We repeat a few iterations to output the final segmentation.

We build the pairwise term to model the local smoothness information. Following the work in (Boykov and Kolmogorov 2004), we construct the pairwise term for a pair of neighboring super-pixels based on color, position and texture information, denoted by subscript C , P and T , respectively. The pairwise term can be written as:

$$\begin{aligned} \varphi_{ij}(y_i, y_j) = & \omega_C \exp\left(-\frac{\|h_C(x_i) - h_C(x_j)\|^2}{2\sigma_C^2}\right) \\ & + \omega_P \exp\left(-\frac{\|h_P(x_i) - h_P(x_j)\|^2}{2\sigma_P^2}\right) \\ & + \omega_T \exp\left(-\frac{\|h_T(x_i) - h_T(x_j)\|^2}{2\sigma_T^2}\right), \end{aligned} \quad (3)$$

where h_C , h_P and h_T are the histogram features for the color, position and texture respectively, ω_C , ω_P and ω_T are the trade-off parameters of different terms. Here σ_C , σ_P and σ_T are the bandwidth parameters. We use the multi-label graph cut to minimize the energy function (Boykov and Kolmogorov 2004).

Iterative refinement of pseudo masks

As shown in Fig. 5, with the resulting pseudo part segmentation masks, we train an FCN based part segmentation model. We use DeepLab (Chen et al. 2016b) segmentation method as our based model with VGG (Simonyan and Zisserman 2014b) as our base network. The trained FCN model is applied to generate the final part segmentation.

We obtain part score maps from the trained FCN model. The part score maps can be incorporated into the unary term in Equation 1 to further improve the pseudo mask generation. In particular, the energy function Equation in 1 can be updated as:

$$E = \sum_i \varphi_i^S(y_i) + \sum_i \varphi_i^N(y_i) + \sum_{i,j} \varphi_{ij}(y_i, y_j), \quad (4)$$

where φ_i^S is same as the unary term from Equation 1, which is constructed based on the skeleton information, $\varphi_i^N(y_i)$ is based on the FCN part score map, and $\varphi_{ij}(y_i, y_j)$ is the same as the pairwise term in Equation 1 to model the local smoothness of neighboring super-pixels.

We generate refined pseudo masks by minimizing the above energy function using graph cut again, and train a new FCN model using the refined pseudo masks. The whole process is illustrated in Fig. 5. In the first iteration, we only use object keypoints to generate pseudo masks and then train the FCN. From the second iteration, we use object keypoints together with FCN part prediction score map from the last iteration to generate new pseudo masks. We repeat these steps for a few iterations. The FCN model in the last iteration is the final model for producing part segmentation prediction.

Table 1: Information of object keypoint annotations in two datasets

	head	torso	arm	leg	object
Human keypoint in PASCAL	Forehead Neck	Neck Shoulder / Hip	Shoulder Elbow / Wrist	Hip Knee / Ankle	All Joint Point
Human keypoint in COCO	Nose Eye / Ear	Neck Shoulder / Hip	Shoulder Elbow / Wrist	Hip Knee / Ankle	All Joint Point

Table 2: Our human part definition in PASCAL VOC Person Part dataset.

head	torso	arm	leg	object
hair / head ear / eye eyebrow mouth neck / nose	torso	lower arm upper arm hand	lower leg upper leg foot	all parts

Correlation network for joint prediction

There is a strong correlation between part and object segmentation. It is expected joint part and object segmentation can benefit each other. In our FCN model, we propose a correlation block to formulate the interaction between parts and objects. Usually, it is easier to segment out an object correctly than segmenting a part. Thus, we propose to use object information to guide part prediction. Fig. 4 shows the framework of the joint inference.

In this module, we generate foreground and background probability for all spatial locations from the object score map. Then we perform element-wise multiplication between the probability map of objects and the part score map to generate a refined part score map. In the training step, we add a dense classification loss to this refined part score map for training. The loss function can be formulated as:

$$L_{part} = \sum_{i=1}^n (-\log(P(Z_i^P | X) \otimes P(Z_i^O | X))), \quad (5)$$

where Z_i^P is the part prediction from the network and Z_i^O is the object prediction. The symbol \otimes indicates element-wise multiplication. Here Z_i^P with $K + 1$ dimensions contains the output of K parts and the background category. Z_i^O contains the foreground probability and the background probability where we repeat the foreground probability K times to match the dimension of Z_i^P .

Experiments

We use the PASCAL VOC Person Part dataset to evaluate our weakly supervised method.

We merge some fine level parts in Person Part dataset to match with our defined part categories based on the keypoint annotations. We focus on four types of human parts: head, torso, arm, and leg. The detailed merge strategy can be found in Table 2.

We use object keypoints for learning our weakly supervised method. Object keypoints can be obtained from PASCAL VOC Human Pose dataset (Xia et al. 2017), or

Table 3: Result comparison (IoU scores) between a fully supervised method with pixel-level annotations and our weakly supervised method with object keypoint annotations.

	Part						Object
	head	torso	arm	leg	bg	mean	
Weakly Supervised (ours only Graph Cut)	48.82	33.41	34.11	32.21	83.81	46.47	52.72
Weakly Supervised (ours VGG)	55.85	35.65	27.97	25.34	87.73	46.50	58.26
Weakly Supervised (ours ResNet)	55.79	40.59	32.63	37.98	87.40	50.86	59.82
Fully Supervised (upper bound)	66.47	47.82	39.93	34.24	92.79	56.25	70.29

from pre-trained object keypoints detector such as Mask RCNN (He et al. 2017) and AlphaPose (Fang et al. 2017; Xiu et al. 2018). If not specifically mentioned, we use the keypoint annotations from the PASCAL VOC Human Pose dataset for training.

Implementation details

In the pseudo mask generation step, we generate the initial confident foreground and background regions from keypoint annotations for constructing the unary item in the graph cut model. For the initial foreground regions, we employ the labeling strategy in Table 1 to generate part labels from the keypoint annotations and keypoint connections. For the initial background regions, we set the regions which are at least 50 pixels away from the nearest keypoint as background regions. The graph-cut optimization step is sensitive to the granularity of the super-pixel. We set the minimum component size as 60.

We trained FCN based segmentation networks using the generated pseudo masks. Our FCN model is based on the DeepLab method (Chen et al. 2016b), and we use VGG-16 (Simonyan and Zisserman 2014a) as our backbone network. We set equal weights for the object loss, the part loss and the refined part loss. We set the batch size to 12 and run 8000 training iterations. We set the learning rate to 0.001 and reduce the learning rate after every 1000 iteration by a factor of 0.5. The momentum is set to 0.9 and the weight decay is set to 0.0005.

Comparison with fully supervised learning

We compare the performance of our weakly supervised learning method with the conventional fully supervised method. As shown in Fig. 1, fully supervised part segmentation methods require expensive pixel-level annotations, while our weakly supervised method only uses object keypoint annotations. The results are shown in Table 3. Our weakly supervised method is able to achieve a comparable

Table 4: Results for using keypoint annotations in the test time. Training column indicates the type of keypoint annotations used for training. Testing column indicates the type of keypoint annotations used in the test time. The segmentation results are improved for incorporating keypoints in the test time using graph cut.

Training	Testing	Part						Object
		head	torso	arm	leg	bg	mean	
Human Pose dataset	Human Pose dataset	58.72	39.89	35.16	33.87	87.86	51.10	60.84
Human Pose dataset	Mask RCNN	56.42	37.87	34.21	30.47	87.80	49.36	58.90
Human Pose dataset	AlphaPose	56.41	38.88	34.03	32.24	87.63	49.84	58.66
Mask RCNN	Mask RCNN	47.79	38.09	33.96	30.24	87.23	47.46	56.78
AlphaPose	AlphaPose	47.25	37.94	33.87	30.37	87.12	47.31	56.09

Table 5: Ablation study of our iterative refinement for pseudo mask generation. The part segmentation results shown below are generated by the FCN based segmentation network trained on the generated pseudo masks. Results are IoU scores.

Iter.	Part						Object
	head	torso	arm	leg	bg	mean	
1	50.95	32.71	26.33	23.53	86.44	43.99	53.00
2	54.35	35.58	27.95	25.71	87.24	46.17	56.70
3	55.19	35.72	28.02	25.58	87.46	46.39	57.38
4	55.70	35.67	27.88	25.62	87.54	46.48	57.68
5	55.85	35.65	27.97	25.34	87.73	46.50	58.26

Table 6: Ablation study of our correlation network for joint learning of objects and parts on PASCAL Human Part dataset. “Part loss only” means the part prediction only uses the part segmentation branch. Results are IoU scores.

	head	torso	arm	leg	bg	mean
Part loss only	49.47	32.19	25.36	22.78	86.08	43.18
Joint learning	50.95	32.71	26.33	23.53	86.44	43.99

result with a fully supervised method, with about 10 % performance drop. Note that the fully supervised method uses the VGG network, same as our VGG based model. Some prediction examples are shown in Fig. 6.

Ablation studies

Iterative refinement. Table 5 shows the IoU scores with different numbers of iterations for the pseudo mask refinement. We can see that our iterative refinement approach effectively improves the segmentation. More iterations remarkably improve the final part segmentation performance and it converges after 3 iterations.

Object-part joint learning. We develop a correlation network for a joint part and object learning. Table 6 shows the results with and without our correlation block, which demonstrates that our correlation block successfully improves the IoU scores for part segmentation.

Using keypoint detectors for learning. We evaluate using different types of keypoint detectors to generate object keypoints for our weakly supervised method including the well-known Mask RCNN based keypoint detection method (He et al. 2017) and AlphaPose method (Fang et al. 2017; Xiu et al. 2018). We report the part segmentation results in terms of IoU in Table 7. It shows that our method is able

Table 7: Results of using different types of keypoint annotations for learning. Results are IoU scores on PASCAL Human Part dataset.

Annotation	Part						Object
	head	torso	arm	leg	bg	mean	
Pose dataset	55.19	35.72	28.02	25.58	87.46	46.39	57.38
Mask RCNN	44.91	34.72	26.49	24.04	86.75	43.38	53.88
AlphaPose	43.87	34.30	27.20	24.83	86.54	43.35	53.51

to incorporate less accurate keypoint annotations generated by keypoint detectors for learning, while still able to achieve very competitive part segmentation results.

Using keypoints in test time. We are able to incorporate the keypoints in the test time to further improve our part segmentation results. We use Graph Cut optimization to jointly consider FCN segmentation results and object keypoints to improve the segmentation results, same as that done in training for the pseudo mark refinement. The results are shown in Table 4. We choose the well known Mask RCNN (He et al. 2017) and AlphaPose (Fang et al. 2017; Xiu et al. 2018) as our keypoint detectors. The results show that using the keypoints in the test time is able to significantly improve the part segmentation performance.

Conclusions

We have proposed a novel weakly supervised human parsing method which only uses object keypoint annotation for learning. Our method significantly reduces human labeling efforts for pixel-level human parsing tasks. Particularly, we have developed an iterative learning approach to generate accurate pseudo masks of parts, and we have also developed a correlation network for joint learning of parts and objects, which improves the part segmentation. Our comprehensive ablation study and performance evaluation have justified the effectiveness and usefulness of the proposed method for human parsing.

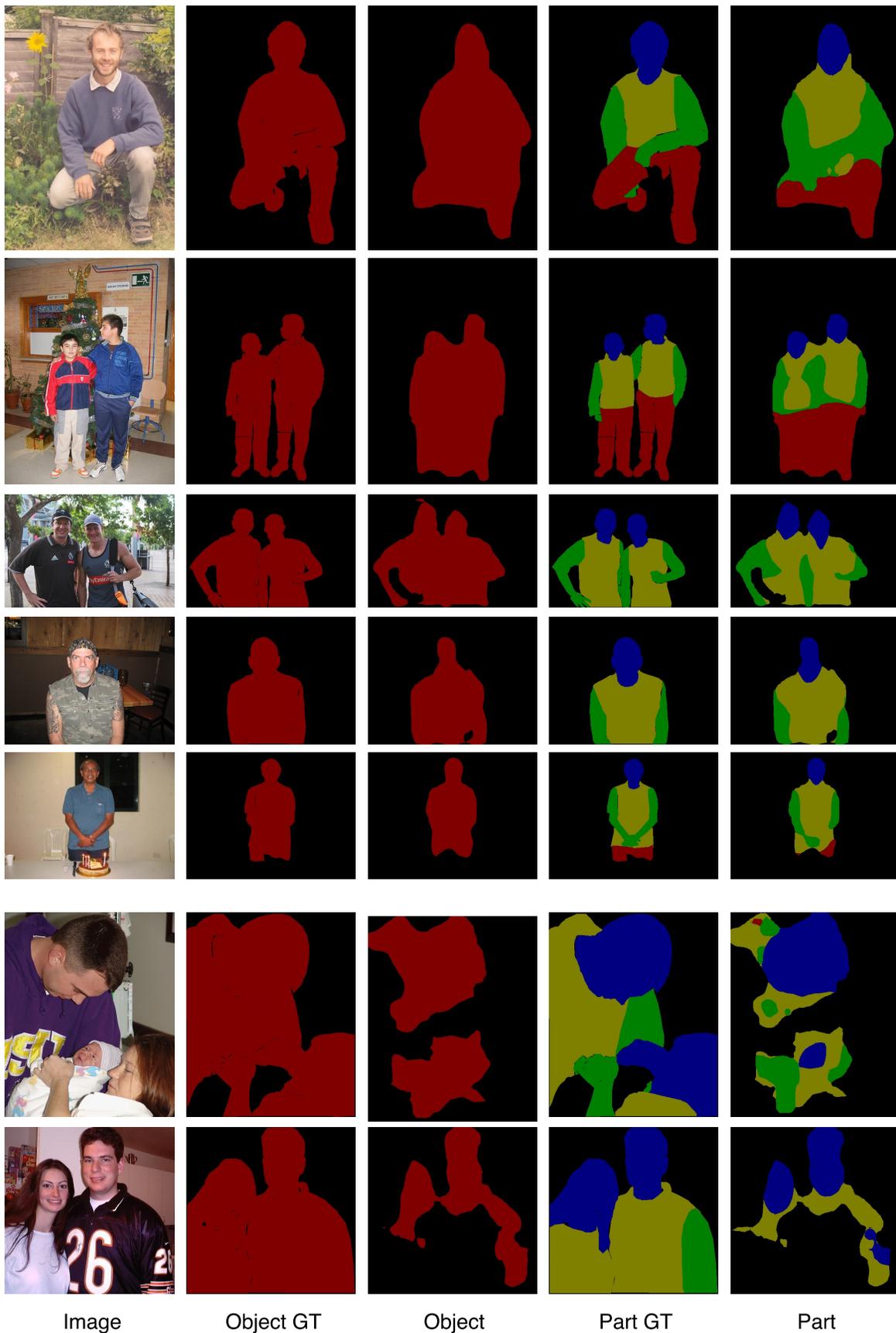


Figure 6: Some examples of our weakly supervised human parsing on PASCAL Human Part dataset. The first five rows show good cases and the last two rows show failure cases.

References

- Bearman, A.; Russakovsky, O.; Ferrari, V.; and Fei-Fei, L. 2016. Whats the point: Semantic segmentation with point supervision. In *European Conference on Computer Vision*, 549–565. Springer.
- Boykov, Y., and Kolmogorov, V. 2004. An experimental comparison of min-cut/max-flow algorithms for energy minimization in vision. *IEEE transactions on pattern analysis and machine intelligence* 26(9):1124–1137.
- Cao, Z.; Simon, T.; Wei, S.-E.; and Sheikh, Y. 2017. Realtime multi-person 2d pose estimation using part affinity fields. In *CVPR*.
- Chen, X.; Mottaghi, R.; Liu, X.; Fidler, S.; Urtasun, R.; and Yuille, A. 2014. Detect what you can: Detecting and representing objects using holistic models and body parts. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.
- Chen, L.-C.; Papandreou, G.; Kokkinos, I.; Murphy, K.; and Yuille, A. L. 2016a. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *arXiv preprint arXiv:1606.00915*.
- Chen, L.-C.; Yang, Y.; Wang, J.; Xu, W.; and Yuille, A. L. 2016b. Attention to scale: Scale-aware semantic image segmentation. In *CVPR*.
- Dai, J.; He, K.; and Sun, J. 2015. Boxesup: Exploiting bounding boxes to supervise convolutional networks for semantic segmentation. In *Proceedings of the IEEE International Conference on Computer Vision*, 1635–1643.
- Fang, H.-S.; Xie, S.; Tai, Y.-W.; and Lu, C. 2017. RMPE: Regional multi-person pose estimation. In *ICCV*.
- He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778.
- He, K.; Gkioxari, G.; Dollár, P.; and Girshick, R. 2017. Mask r-cnn. In *Computer Vision (ICCV), 2017 IEEE International Conference on*, 2980–2988. IEEE.
- Li, J.; Zhao, J.; Wei, Y.; Lang, C.; Li, Y.; and Feng, J. 2017. Towards real world human parsing: Multiple-human parsing in the wild. *arXiv preprint arXiv:1705.07206*.
- Liang, X.; Liu, S.; Shen, X.; Yang, J.; Liu, L.; Dong, J.; Lin, L.; and Yan, S. 2015. Deep human parsing with active template regression. *IEEE transactions on pattern analysis and machine intelligence* 37(12):2402–2414.
- Lin, T.-Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; and Zitnick, C. L. 2014. Microsoft coco: Common objects in context. In *European conference on computer vision*, 740–755. Springer.
- Lin, D.; Dai, J.; Jia, J.; He, K.; and Sun, J. 2016. Scribble-sup: Scribble-supervised convolutional networks for semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 3159–3167.
- Lin, G.; Milan, A.; Shen, C.; and Reid, I. 2017. RefineNet: Multi-path refinement networks for high-resolution semantic segmentation. In *CVPR*.
- Long, J.; Shelhamer, E.; and Darrell, T. 2015. Fully convolutional models for semantic segmentation. In *CVPR*, volume 3, 4.
- Papandreou, G.; Chen, L.-C.; Murphy, K.; and Yuille, A. L. 2015. Weakly- and semi-supervised learning of a dcnn for semantic image segmentation. In *ICCV*.
- Pathak, D.; Shelhamer, E.; Long, J.; and Darrell, T. 2014. Fully convolutional multi-class multiple instance learning. *arXiv preprint arXiv:1412.7144*.
- Shen, T.; Lin, G.; Liu, L.; Shen, C.; and Reid, I. D. 2017. Weakly supervised semantic segmentation based on co-segmentation. In *BMVC*.
- Simonyan, K., and Zisserman, A. 2014a. Very deep convolutional networks for large-scale image recognition. *CoRR* abs/1409.1556.
- Simonyan, K., and Zisserman, A. 2014b. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.
- Tsogkas, S.; Kokkinos, I.; Papandreou, G.; and Vedaldi, A. 2015. Deep learning for semantic part segmentation with high-level guidance. *arXiv preprint arXiv:1505.02438*.
- Uijlings, J. R.; Van De Sande, K. E.; Gevers, T.; and Smeulders, A. W. 2013. Selective search for object recognition. *International journal of computer vision* 104(2):154–171.
- Vernaza, P., and Chandraker, M. 2017. Learning random-walk label propagation for weakly-supervised semantic segmentation. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 3.
- Wang, P.; Shen, X.; Lin, Z.; Cohen, S.; Price, B.; and Yuille, A. 2015. Joint object and part segmentation using deep learned potentials. *ICCV*.
- Xia, F.; Wang, P.; Chen, L.-C.; and Yuille, A. L. 2016. Zoom better to see clearer: Human and object parsing with hierarchical auto-zoom net. In *European Conference on Computer Vision*, 648–663. Springer.
- Xia, F.; Wang, P.; Chen, X.; and Yuille, A. 2017. Joint multi-person pose estimation and semantic part segmentation. *arXiv preprint arXiv:1708.03383*.
- Xiu, Y.; Li, J.; Wang, H.; Fang, Y.; and Lu, C. 2018. Pose Flow: Efficient online pose tracking. *ArXiv e-prints*.
- Zhang, T.; Lin, G.; Cai, J.; Shen, T.; Shen, C.; and Kot, A. C. 2018. Decoupled Spatial Neural Attention for Weakly Supervised Semantic Segmentation. *ArXiv e-prints*.