arXiv:1803.10683v1 [cs.CV] 28 Mar 2018

Pose2Seg: Human Instance Segmentation Without Detection

Ruilong Li, Xin Dong, Zixi Cai, Dingcheng Yang, Haozhi Huang[1], Song-Hai Zhang, Paul Rosin[2], Shi-Min Hu

Tsinghua University [1]Tencent AI Lab [2]Cardiff University

Abstract. The general method of image instance segmentation is to perform the object detection first, and then segment the object from the detection bounding-box. More recently, deep learning methods like Mask R-CNN [1] perform them jointly. However, little research takes into account the uniqueness of the "human" category, which can be well defined by the pose skeleton. In this paper, we present a brand new pose-based instance segmentation framework for humans which separates instances based on human pose, not proposal region detection. We demonstrate that our pose-based framework can achieve similar accuracy to the detection-based approach, and can moreover better handle occlusion, which is the most challenging problem in the detection-based framework.

Keywords: instance segmentation, human pose

1 Introduction

In recent years, research related to "humans" in the computer vision community has become increasingly active because of the high demand for real-life applications. There has been much good research in the fields of human pose estimation, pedestrian detection, portrait segmentation, and face keypoint detection, much of which has already produced practical value in real life. This paper focuses on multi-person pose estimation and human instance segmentation, and proposes a pose-based human instance segmentation framework.

Multi-Person Pose Estimation and *General Objects Instance Segmentation* are both challenging problems. The aim of multi-person pose estimation is to estimate the pose of each human instance in the image, which is defined by a series of body parts like head, shoulders, hands, feet and so on. General object instance segmentation aims to predict pixel-level labels for each object instance in the image. The solutions to those two problems need a common process, which is to detect and isolate each instance. This process can be considered as an independent challenging problem called *Object Detection*. Because of the loss of depth information in the two-dimensional image, it is difficult to separate objects of the same category which are substantially overlapping. There has been a lot of powerful baseline systems for object detection, such as Fast/Faster R-CNN [2, 3], YOLO [4], which mostly follow a basic rule: first generate a large number of proposal regions, then remove the redundant regions using *Non-maximum*

Fig. 1. Basic idea of bottom-up multi-person pose estimation. First detect body part keypoints of all the people, and then group or connect those parts to form several instances of human pose.

Fig. 2. The large overlap problem in *object detection* methods can be improved using human poses.

Suppression (NMS). However, when two objects of the same category have a large overlap, NMS always treats one of them as a redundant proposal region and eliminates it. This means that almost all the object detection methods cannot deal with the situation of large overlaps.

Although object detection methods are widely used by many multi-person pose estimation frameworks, some powerful bottom-up methods [5, 6] which do not rely on object detection also achieved good performance, including the *COCO keypoints challenge 2016 winner* [5]. The main idea of the bottom-up methods is to first detect keypoints for each body part for all the people, and then group or connect those parts to form several instances of human pose, as shown in Figure 1. There are several advantages of this kind of bottom-up method. First, the run time cost will not increase with an increased number of people in the image. Secondly, which is the most important one, two intertwined human instances with a large overlap can be separated when connecting their body parts. The large overlap problem in object detection methods can be improved using human poses, as shown in Figure 2. "Human" is a special category in the visual community. Unlike "table", "bed" and so on, human pose has a perfect definition, which can be used to locate and separate different instances in the image. So we present a brand new pose-based instance segmentation framework for humans which separates instances based on human pose rather than region proposal detection. Our pose-based framework works seamlessly with existing

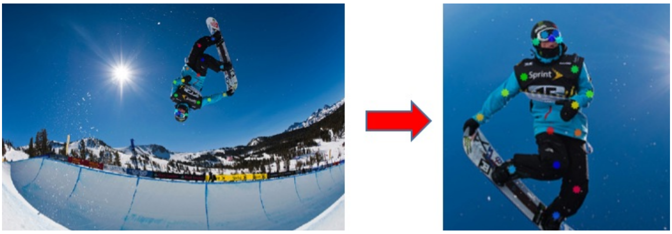

Fig. 3. An example of correcting objects with strange pose to a standard pose.

bottom-up pose estimation methods, and works better than the detection-based framework in the case of occlusion.

Generally, there is an align module in the instance segmentation framework, for example, *ROI-Align* in Mask R-CNN. The align module is used to crop the objects from the image using detection bounding boxes, and resize the objects to a uniform scale. Because of the receptive field of the neural network, it is easier to learn using a uniform scale rather than a variety of scales. Since it is hard to find a bounding box accurately from the object using human pose, we proposed an align module based on human pose, called *AffineAlign*. First, we automatically select some "standard" poses from the training set as templates, as shown in Figure 5. Then, for each estimated human pose, we find the most similar template and calculate an affine transformation matrix between them by solving an optimization problem, as mentioned in Section 3.2. Finally, the affine matrix is applied to the feature map, and the human instance is aligned to the center with a uniform scale, as shown in Figure 5. But unlike bounding-box based align algorithms such as ROI-Align, the AffineAlign operation is a combination of scale, translation and rotation. So an extra advantage of using AffineAlign is that we can correct some objects with strange poses to a standard pose, like the inverted skiing human in Figure 3. A standard pose alignment makes the segmentation module easier to learn, and leads to a better accuracy.

Additionally, the human pose and human mask are not independent. Human pose can be approximately considered as a skeleton of the mask of the human instance. So we explicitly use human pose to guide the segmentation module by concatenating the pose feature to the instance feature map after Affine-Align. Our experiment shows it can further improve the segmentation accuracy.

Our main contributions can be summarized as follows:

- We propose a brand new pose-based human instance segmentation framework which works better than the detection-based framework in cases with occlusion.
- We propose a pose-based align module, called *AffineAlign*, which can align image windows into a uniform scale and direction based on human pose.
- We explicitly use human pose to guide the segmentation module and achieve a further improvement of the segmentation accuracy.

2 Related Work

2.1 Multi-Person Pose Estimation

Human pose estimation is a classic and fundamental work in computer vision. In recent years, with the rapid development of convolutional neural networks, both the precision and efficiency of human pose estimation have been greatly improved. Single-person pose estimation is the fundamental element of multi-person pose estimation, and a lot of excellent work [7–11] has made great success on improving the performance of single-person pose estimation. In comparison, multi-person pose estimation is a more challenging problem because of the unknown number of people, variety of scales of people and unpredictable interactions between several people. Prior and concurrent works about multi-person pose estimation can be roughly divided into two categories: top-down methods and bottom-up methods.

Top-down methods Top-down methods [1, 12–15] first employ human detection to crop each person, and then use a single-person pose estimation method on each human instance. Currently the state-of-art result of multi-person pose estimation based on the COCO [16] dataset is using top-down methods. Papandreu et al. [12] propose a simple but powerful two-stage pipeline for the problem, first using Faster-RCNN to generate human bounding boxes, and then using Resnet [17] for pose estimation. Fang et al. [13] propose the regional multi-person pose estimation (RMPE) framework to reduce the negative impact of inaccurate bounding boxes. Huang et al. [14] and Chen et al. [15] employ a cascaded pyramid network to generate coarse/fine keypoints results respectively.

Bottom-up methods Bottom-up methods [5,6,18,19] first detect body part keypoints of all the people, and then cluster these parts into instances of human pose, which involves complex post-processing. Pishchulin et al. [18] propose a complex framework of partitioning and labeling body-parts generated using a CNN. They solve the problem as an integer linear program, and jointly generate the detection and pose estimation results. Insafutdinov et al [19] use Resnet [17] to improve precision, and propose image-conditioned pairwise terms to increase speed. Cao et al. [5] use knowledge of the human structure, and predict a keypoints heatmap and PAFs, and finally combine the body parts. Newell et al. [6] propose a framework using associative embedding for joint detection and grouping. They designed a tag score map for each body part and propose an embedding loss for the tag score map. They use the score map to group body part keypoints and achieved performance that was state-of-art amongst bottom-up methods for multi-person pose on the MPII [20] and COCO [16] datasets.

2.2 Instance Segmentation

With the development of deep convolutional neural networks in recent years, semantic segmentation [21–25] and instance segmentation [26–33] have made significant progress. Due to the difficult of segmenting the same kind of objects, instance segmentation is more challenging. Most instance segmentation

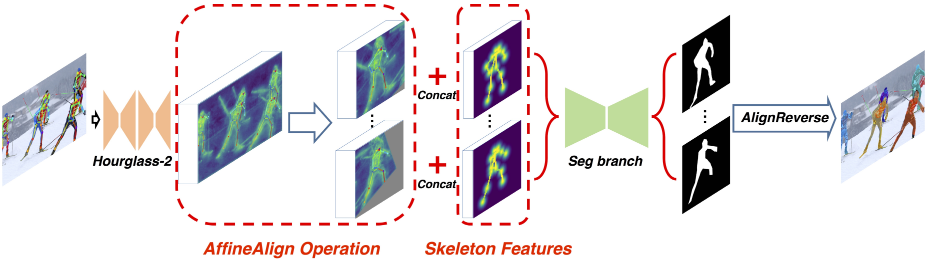

Fig. 4. Our network structure. We use an AffineAlign operation to align the image windows based on human pose. And then we gennerate skeleton features for each human instance to help the segmentation branch concentrate. Finally, we reverse the AffineAlign operation and form the final instance segmentaion result.

frameworks combine detection and semantic segmentation to generate instance results. Some works [26–29] employ a multi-stage pipeline which first uses detection to generate bounding boxes and then applies semantic segmentation. Others [30–32] employ a tighter integration of detection and segmentation, e.g. [30] jointly and simultaneously detects and segments in an end-to-end framework. Mask R-CNN [1] is the state-of-art framework on the COCO [16] dataset competition. Unlike the above methods which rely on object detection, our pipeline employs bottom-up human pose estimation to replace object detection in the framework.

2.3 Harness Human pose estimation for Instance Segmentation

There are two typical works that combine human pose estimation and instance segmentation. The Mask RCNN [1] approach detects objects while generating instance segmentation and human pose estimation simultaneously in a single framework. But in their work, Mask R-CNN [1] with mask-only performs better than combining keypoints and masks in the instance segmentation task. Pose2Instance [34] proposes a cascade network to harness human pose estimation for instance segmentation. Both the two works rely on human detection, and perform poorly when two bounding boxes have a large overlap. Our approach uses human pose estimation to replace bounding boxes and uses keypoints to improve the precision of instance segmentation.

3 Method

3.1 Overview

We proposed an end-to-end human instance segmentation framework based on human pose estimation. Our overall structure is shown in Figure 4, which takes both the image and the human pose estimation results as the input. The base network we use is *stack hourglass* [11] because of its advantages in global feature

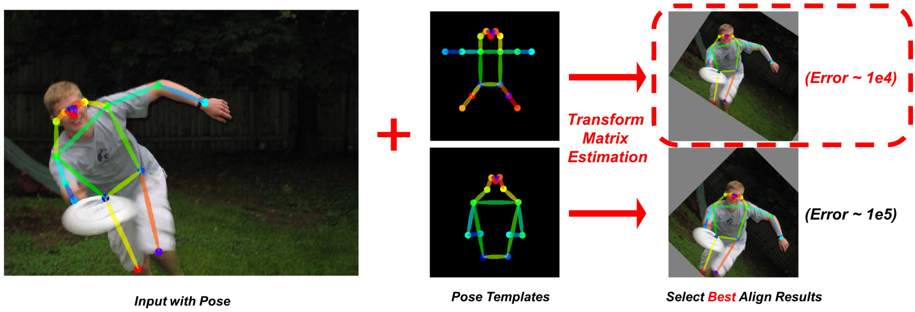

Fig. 5. Our AffineAlign Operation. We first estimate the affine transformation matrix H between human pose and the templates, and then chose H with the smallest error.

extraction. Then we use an align module to align *regions of interest (ROIs)* to a uniform size, which is 64×64 in this paper, based on the human pose estimation results, called *AffineAlign*. In the meantime, we generate skeleton features for each human instance and concatenate them to the *ROIs*. Our experiments show that explicitly adding the skeleton information to the network provides better information in the segmentation branch. Our segmentation branch is simply 10 repeated units of Resnet [17] to achieve a global receptive field on the ROIs. Finally, a reverse affine align operation is applied to get the final segmentation results. We describes each of those steps in the following subsections.

3.2 AffineAlign

The variety of scales and complex poses of humans makes human instance segmentation a challenging problem. Our *AffineAlign* operation is inspired by the ROIPooling in Faster R-CNN [3] and ROIAlign in Mask R-CNN [1]. But unlike them, we align the people based on human pose instead of bounding-boxes. With the information provided by human pose, the AffineAlign Operation can straighten people who are in strange poses, and separate people with large overlaps. Specifically, as shown in Figure 5, we use a set of pose templates to represent the standard poses in the dataset. For each pose detected in the image, we first estimate the affine transformation matrix H between it and the templates, and then chose H with the smallest error. The operations in H contain rotation, scale, translation and left-right reflection. We do not include shear in H because distortions are not good for the square-kernel convolution neural network. Finally, we apply H to the image or features and transform it to the desired resolution.

Human Pose Representation Human poses are represented as a list of vectors. Let vector $P = (C_1, C_2, ..., C_n) \in \mathbb{R}^{n\times 2}$ represent the pose of a single person, where $C_i = (x, y) \in \mathbb{R}^2$ is a 2D vector representing the coordinates of a single part (such as right-shoulder, left-ankle) and n is a dataset related param-

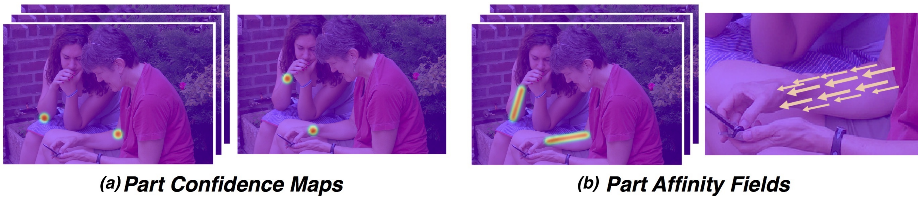

Fig. 6. Our Skeleton Features, including (a) confidence maps for body part, and (b) part affinity fields (PAFs) for skeletons. The last image in (b) shows a zoom-in view of the PAFs.

eter meaning the total number of parts in a single pose, which is 17 in COCO. If some parts are missing from the image, we set $C_i = (0, 0)$.

Pose Templates Human poses have a huge variety, and cannot be represented by a single pose template. And artificially defined template can not represent the distribution of the dataset. To learn the distribution of poses in the dataset we cluster the pose templates from the training set. We use k-means clustering to cluster the poses in the training set into K clusters (experimentally, we choose K=2 in this paper for the COCO dataset). Then we use the poses from the center of clusters to represent the whole group and form a set of pose templates, as shown in Figure 5. We find that the most frequent pose is a standard full-body pose, and the second one is a half-body pose. Although a single full-body template is enough sufficient to align all kinds of different poses, obviously, a half-body pose template can achieve a better alignment performance on those half-body images. So finally, we use both full-body pose and half-body pose as our pose templates. More pose templates are not necessary.

Estimate Affine Transformation Matrix Let vector $P_t = (C_{t1}, C_{t2}, ..., C_{tn})$ to represent a pose template, and P_d represent a single person pose estimation result. We optimize Eq. 1 to estimate an affine transformation matrix H which makes the pose coordinates are as near as possible to the template coordinates after the affine transformation. Since we have two templates, we use the optimized error value, calculated by Eq. 2, to choose the best template for each estimated pose, as shown in Figure 5.

$$H^* = \underset{H}{\arg\min} \ \|H \cdot P_d - P_t\|. \tag{1}$$

$$Error = \|H^* \cdot P_d - P_t\| \tag{2}$$

3.3 Skeleton features

Human pose can be approximately considered as a skeleton of the mask of the human instance. So an intuitive idea to further improve the accuracy of segmentation is to use the information available from the human skeleton. Therefore, we generate a 44-channel feature for each human instance, called Skeleton Features, and concatenate these features to the *ROIs* of image features, as shown in Figure 4.

Fig. 7. Samples from COCOHUMAN-OC. Some "*occlusion*" objects from COCO2017 validation set.

Figure 6 shows our Skeleton Features. We adopt the part affinity fields (PAFs) from [5], which is a 2-channel vector field map for each skeleton. We use PAFs to represent the skeleton structure of a human pose. With 11 skeletons defined in the COCO dataset, PAFs is a 22-channel feature map for each human pose instance. We also use part confidence maps for body parts to emphasize the importance of those regions around the body part keypoints. In general cases, there is a high probability that these regions belongs to the mask of this person. For the COCO dataset, each human pose has a 17-channel part confidence map and a 22-channel PAFs map. So the total number of channels in our Skeleton Features is 39 for each human instance.

4 Experiments

4.1 Dataset

COCOHUMAN As far as we know, there are few public datasets which have labels for both human pose and human instance segmentation. COCO is the largest dataset that meets both of these requirements. The COCO dataset has instance segmentation labels for 81 categories of objects including human-object. Objects in the dataset are divided into three groups by *Area* (the number of pixels in the segmentation mask). Those with $Area < 32^2$ are defined as *Small* objects. Those among 32^2 and 96^2 are defined as *Medium* objects. Others with $Area > 96^2$ are defined as *Large* objects. For most of the *Medium* and *Large* objects in the human category, COCO has annotations for its pose keypoints. So we select this subset of COCO as our experiment dataset, with both human pose and human instance segmentation labels. Our trainset/valset is a subset of the COCO2017 trainset/valset, which contains 56599/2346 images with

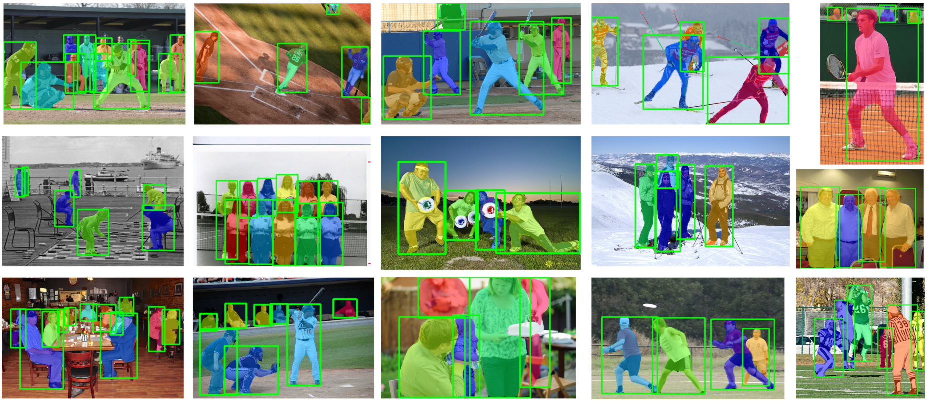

Fig. 8. More results of our instance segmentation framework. Bounding-boxes are generated using predicted masks for better visualization.

149813/6352 instance annotations of human-object. We call this subset *COCO-HUMAN* dataset.

COCOHUMAN-OC To evaluate our method's capacity for handling occlusion cases, we first define "emphocclusion" in the COCOHUMAN dataset. Occlusion objects usually have a large overlap with other objects. So we consider objects with $IOU > 0.5$ as a pair of "emphocclusion" objects. We choose those occlusion cases from the COCOHUMAN validation set, and form a new dataset, called COCOHUMAN-OC. This dataset contains 44 images with 88 instance annotations, all containing large overlaps with others. Some samples are shown in Figure 7.

4.2 Performance

vs. Mask-RCNN on COCOHUMAN In this experiment, our framework is trained end-to-end on the COCOHUMAN training set with annotations of pose keypoints and segmentation masks. No bounding-box annotations were used. We evaluate our model on the COCOHUMAN validation set using groundtruth keypoints as input, and get an average precision of 47.7% on the instance segmentation task. We also evaluate the performance of our model under the predicted pose keypoints using a bottom-up method [6], and achieve an average of 45.6%. This indicates that our pose-based approach works seamlessly with existing pose estimation methods. We compared our methods with Mask-RCNN [1], the well known detection based instance segmentation framework. For Mask-RCNN [1], we use the author's released code and model from [35], which were trained using the whole COCO training set, and evaluate the model on the COCOHUMAN validation set. As shown in Table 1, our framework with groundtruth (Gt) keypoints achieves equal performance with Mask-RCNN. This demonstrates that the pose-based framework can achieve similar performance to the state-of-art

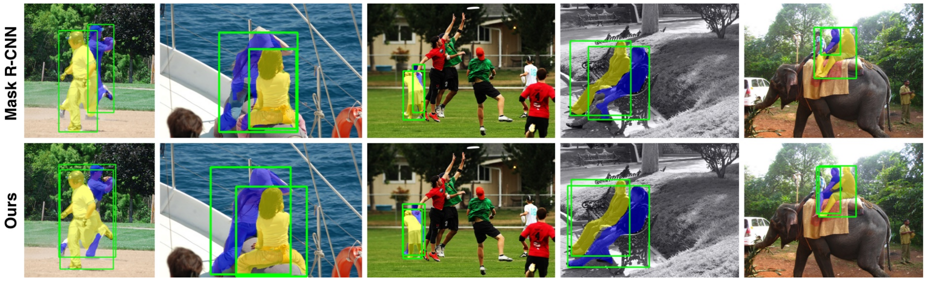

Fig. 9. Our method's results vs. Mask R-CNN [1] on occlusion cases. Bounding-boxes in our results are generated using predicted masks for better visualization and comparison.

detection based framework on the task of instance segmentation. Figure 8 and Figure 10 show the results of the instance segmentation framework and our AffineAlign operation, respectively.

Table 1. Performance on COCOHUMAN validation set. "Pt Keypoints" refers to using keypoints predicted by a bottom-up method [6].

Methods	Backbone	AP (Medium+Large)
Mask-RCNN	Resnet50-C4	0.477
Our (Gt Keypoints)	Hourglass-2	0.477
Our (Pt Keypoints)	Hourglass-2	0.456

vs. Mask-RCNN on COCOHUMAN-OC We evaluate our method's capacity for handling occlusion cases compared with Mask-RCNN [1] on the COCOHUMAN-OC validation set. Table 2 shows our pose-based framework works better than the detection based framework on the occlusion cases. Some results are shown in Figure 9. Because of NMS, detection based frameworks like Mask R-CNN [1] cannot well handle objects with large overlaps. Even though instance objects can be separated, there are still some parts missing due to the threshold setting in NMS. But this is not a limitation in our pose-based framework. We can achieve a full-body segmentation result for each instance even though there is a serious occlusion problem, as shown in Figure 9.

Table 2. Performance on COCOHUMAN-OC validation set. "Pt Keypoints" meaning using keypoints predicted by a bottom-up method [6].

Methods	AP	AP (medium)	AP (large)
Mask-RCNN (End-to-End)	0.055	0.033	0.095
Our (Gt Keypoints)	0.141	0.072	0.273
Our (Pt Keypoints)	0.073	0.037	0.127

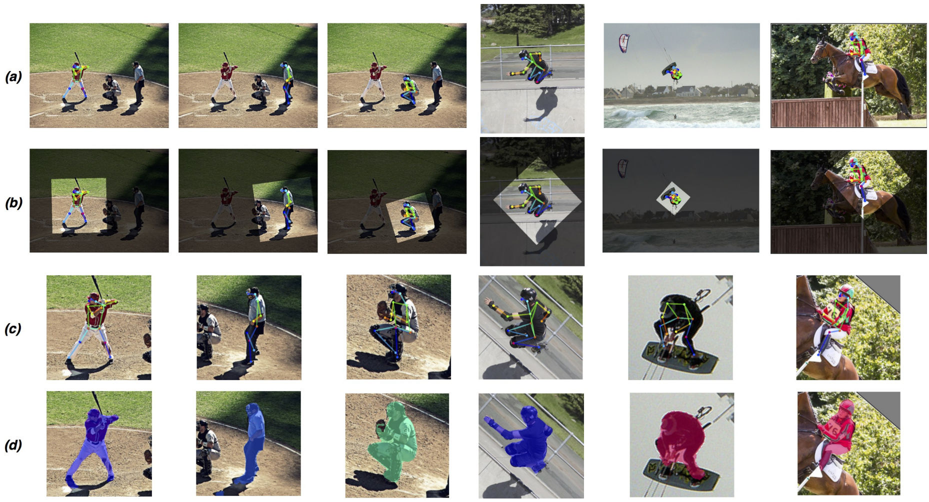

Fig. 10. More results of our AffineAlign Operation. AffineAlign takes human pose as input to align the image windows to a uniform size and direction. (a) shows the inputs, (b) shows to align window on the original image, (c) is the align results of AffineAlign Operation, (d) is the segmentation results of our framework.

5 Conclusion

In this paper, we propose a pose-based human instance segmentation framework. We designed an AffineAlign Operation for selecting Regions Of Interest (ROIs) based on pose instead of bounding-boxes. We explicitly concatenate the human pose skeleton feature to the image feature in the network. Compared to the traditional detection based instance segmentation framework, our pose-based system can achieve a similar performance in the general case, and has better performance in occlusion cases.

References

1. He, K., Gkioxari, G., Dollr, P., Girshick, R.: Mask R-CNN. (2017)
2. Girshick, R.: Fast r-cnn. In: Computer Vision (ICCV), 2015 IEEE International Conference on, IEEE (2015) 1440–1448
3. Ren, S., He, K., Girshick, R., Sun, J.: Faster R-CNN: Towards real-time object detection with region proposal networks. In: Advances in neural information processing systems. (2015) 91–99
4. Redmon, J., Divvala, S., Girshick, R., Farhadi, A.: You only look once: Unified, real-time object detection. In: Proceedings of the IEEE conference on computer vision and pattern recognition. (2016) 779–788
5. Cao, Z., Simon, T., Wei, S.E., Sheikh, Y.: Realtime multi-person 2D pose estimation using part affinity fields. In: CVPR. (2017)
6. Newell, A., Huang, Z., Deng, J.: Associative embedding: End-to-end learning for joint detection and grouping. (2016)
7. Yang, W., Li, S., Ouyang, W., Li, H., Wang, X.: Learning feature pyramids for human pose estimation. In: The IEEE International Conference on Computer Vision (ICCV). Volume 2. (2017)
8. Chen, Y., Shen, C., Wei, X.S., Liu, L., Yang, J.: Adversarial PoseNet: A structure-aware convolutional network for human pose estimation. CoRR, abs/1705.00389 **2** (2017)
9. Chu, X., Yang, W., Ouyang, W., Ma, C., Yuille, A.L., Wang, X.: Multi-context attention for human pose estimation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2017) 1831–1840
10. Ning, G., Zhang, Z., He, Z.: Knowledge-guided deep fractal neural networks for human pose estimation. IEEE Transactions on Multimedia (2017)
11. Newell, A., Yang, K., Deng, J.: Stacked hourglass networks for human pose estimation. In: European Conference on Computer Vision, Springer (2016) 483–499
12. Papandreou, G., Zhu, T., Kanazawa, N., Toshev, A., Tompson, J., Bregler, C., Murphy, K.: Towards accurate multi-person pose estimation in the wild. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2017) 4903–4911
13. Fang, H.S., Xie, S., Tai, Y.W., Lu, C.: RMPE: Regional multi-person pose estimation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2017) 2334–2343
14. Huang, S., Gong, M., Tao, D.: A coarse-fine network for keypoint localization. In: The IEEE International Conference on Computer Vision (ICCV). Volume 2. (2017)
15. Chen, Y., Wang, Z., Peng, Y., Zhang, Z., Yu, G., Sun, J.: Cascaded pyramid network for multi-person pose estimation. arXiv preprint arXiv:1711.07319 (2017)
16. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: European conference on computer vision, Springer (2014) 740–755
17. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. (2016) 770–778
18. Rajchl, M., Lee, M.C., Oktay, O., Kamnitsas, K., Passerat-Palmbach, J., Bai, W., Damodaram, M., Rutherford, M.A., Hajnal, J.V., Kainz, B., et al.: Deepcut: Object segmentation from bounding box annotations using convolutional neural networks. IEEE Transactions on Medical Imaging **36**(2) (2017) 674–683

19. Insafutdinov, E., Pishchulin, L., Andres, B., Andriluka, M., Schiele, B.: DeeperCut: A deeper, stronger, and faster multi-person pose estimation model. In: European Conference on Computer Vision, Springer (2016) 34–50
20. Andriluka, M., Pishchulin, L., Gehler, P., Schiele, B.: 2d human pose estimation: New benchmark and state of the art analysis. In: Proceedings of the IEEE Conference on computer Vision and Pattern Recognition. (2014) 3686–3693
21. Long, J., Shelhamer, E., Darrell, T.: Fully convolutional networks for semantic segmentation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. (2015) 3431–3440
22. Zheng, S., Jayasumana, S., Romera-Paredes, B., Vineet, V., Su, Z., Du, D., Huang, C., Torr, P.H.: Conditional random fields as recurrent neural networks. In: Proceedings of the IEEE International Conference on Computer Vision. (2015) 1529–1537
23. Chen, L.C., Papandreou, G., Kokkinos, I., Murphy, K., Yuille, A.L.: Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. IEEE Transactions on Pattern Analysis and Machine Intelligence (2017)
24. Badrinarayanan, V., Kendall, A., Cipolla, R.: Segnet: A deep convolutional encoder-decoder architecture for image segmentation. IEEE Transactions on Pattern Analysis and Machine Intelligence **39**(12) (2017) 2481–2495
25. Chen, L.C., Papandreou, G., Schroff, F., Adam, H.: Rethinking atrous convolution for semantic image segmentation. arXiv preprint arXiv:1706.05587 (2017)
26. Hariharan, B., Arbeláez, P., Girshick, R., Malik, J.: Simultaneous detection and segmentation. In: European Conference on Computer Vision, Springer (2014) 297–312
27. Hariharan, B., Arbeláez, P., Girshick, R., Malik, J.: Hypercolumns for object segmentation and fine-grained localization. In: Proceedings of the IEEE conference on computer vision and pattern recognition. (2015) 447–456
28. Girshick, R., Iandola, F., Darrell, T., Malik, J.: Deformable part models are convolutional neural networks. In: Proceedings of the IEEE conference on Computer Vision and Pattern Recognition. (2015) 437–446
29. Dai, J., He, K., Sun, J.: Convolutional feature masking for joint object and stuff segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2015) 3992–4000
30. Li, Y., Qi, H., Dai, J., Ji, X., Wei, Y.: Fully convolutional instance-aware semantic segmentation. In: IEEE Conf. on Computer Vision and Pattern Recognition (CVPR). (2017) 2359–2367
31. Pinheiro, P.O., Collobert, R., Dollár, P.: Learning to segment object candidates. In: Advances in Neural Information Processing Systems. (2015) 1990–1998
32. Dai, J., He, K., Li, Y., Ren, S., Sun, J.: Instance-sensitive fully convolutional networks. In: European Conference on Computer Vision, Springer (2016) 534–549
33. Liu, S., Jia, J., Fidler, S., Urtasun, R.: Sgn: Sequential grouping networks for instance segmentation. In: The IEEE International Conference on Computer Vision (ICCV). (2017)
34. Tripathi, S., Collins, M., Brown, M., Belongie, S.: Pose2instance: Harnessing keypoints for person instance segmentation. arXiv preprint arXiv:1704.01152 (2017)
35. Girshick, R., Radosavovic, I., Gkioxari, G., Dollár, P., He, K.: Detectron. `https://github.com/facebookresearch/detectron` (2018)