Pose2Seg: Detection Free Human Instance Segmentation

Song-Hai Zhang[1,2], Ruilong Li[1,2], Xin Dong[1], Paul Rosin[3], Zixi Cai[1], Xi Han[1], Dingcheng Yang[1], Haozhi Huang[4] and Shi-Min Hu[1,2]

[1] Tsinghua University [2] BNRist [3] Cardiff University [4] Tencent AI Lab

{shz, shimin}@tsinghua.edu.cn, {li-rl16, dong-x16, caizx15, x-han15, ydc15}@mails.tsinghua.edu.cn, RosinPL@cardiff.ac.uk, matthzhuang@tencent.com

Abstract

The standard approach to 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 [14] perform them jointly. However, little research takes into account the uniqueness of the "human" category, which can be well defined by the pose skeleton. Moreover, the human pose skeleton can be used to better distinguish instances with heavy occlusion than using bounding-boxes. In this paper, we present a brand new pose-based instance segmentation framework[1] for humans which separates instances based on human pose, rather than proposal region detection. We demonstrate that our pose-based framework can achieve better accuracy than the state-of-art detection-based approach on the human instance segmentation problem, and can moreover better handle occlusion. Furthermore, there are few public datasets containing many heavily occluded humans along with comprehensive annotations, which makes this a challenging problem seldom noticed by researchers. Therefore, in this paper we introduce a new benchmark "Occluded Human (OCHuman)"[2], which focuses on occluded humans with comprehensive annotations including bounding-box, human pose and instance masks. This dataset contains 8110 detailed annotated human instances within 4731 images. With an average 0.67 MaxIoU for each person, OCHuman is the most complex and challenging dataset related to human instance segmentation. Through this dataset, we want to emphasize occlusion as a challenging problem for researchers to study.

[1]Codes are available: https://github.com/liruilong940607/Pose2Seg
[2]Dataset is available: https://github.com/liruilong940607/OCHumanApi

Figure 1: Heavily occluded people are better separated using *human pose* than using *bounding-box*.

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 [1, 2, 6, 14, 20, 26, 40], pedestrian detection [25, 41, 42], portrait segmentation [35, 36, 37], and face recognition [18, 23, 24, 27, 39, 43, 44], 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.

General Object Instance Segmentation is a challenging problem which aims to predict pixel-level labels for each object instance in the image. Currently, those instance segmentation methods with highest accuracy [3, 14, 19, 30] are all based on powerful *object detection* baseline methods, such as Fast/Faster R-CNN [9, 33], YOLO [32], which mostly follow a basic rule: first generate a large number of proposal regions, then remove the redundant regions using *Non-maximum Suppression (NMS)*. However, when two objects of the same category have a large overlap, NMS will treat 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. Moreover, even if the detection methods sometimes successfully detect two instances, the bounding-box is not suitable for instance segmentation in occluded cases. If two in-

stances are heavily intertwined, they will both appear in the same bounding-box (like the case in Figure 1), which makes it hard for the segmentation network to identify which instance should be the target in this *Region of Interest (RoI)*.

However, "human" is a special category in the computer vision community, and can be well defined by the pose skeleton. As shown in Figure 1, Human pose skeletons are more suitable for distinguishing two heavily intertwined people, because they can provide more distinct information about a person than bounding-boxes, such as the location and visibility of different body parts. *Multi-Person Pose Estimation* is also a very active topic in recent years, and there is already good progress [1, 2, 6, 16, 26, 29] on tackling this problem. Although object detection methods are widely used by many multi-person pose estimation frameworks, some powerful bottom-up methods [1, 26] which do not rely on object detection also achieved good performance, including the *COCO keypoints challenge 2016 winner* [1]. 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, which makes it possible to seperate two intertwined human instances with a large overlap. Based on this observation, we present a 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 bottom-up pose estimation methods, and works better than the detection-based framework, especially 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. 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 *Affine-Align*, which is a combination of scale, translation, rotation and left-right flip. An extra advantage of using *Affine-Align* is that we can correct some objects with strange poses to a standard pose, like the inverted skiing human in Figure 2.

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 *Skeleton* features to the instance feature map after *Affine-Align*. Our experiments demonstrate our *Skeleton* features not only help to improve the accuracy of segmentation, but also give our network the ability to easily distinguish different instances that are heavily intertwined in the same RoI.

Severe occlusion between human bodies is often encountered in life, but current human-related public datasets either do not contain many severe occlusion situations [5, 8, 21], or lack comprehensive annotations of the human instances [34]. Therefore, we introduce a new benchmark *"Occluded Human (OCHuman)"* in this paper, which focuses on heavily occluded humans with comprehensive annotations including bounding-boxes, human poses and instance masks. This dataset contains 8110 detailed annotated human instances within 4731 images. On average, over 67% of the bounding-box area of a human is occluded by one or several other persons, which makes this dataset the most complex and challenging dataset related to humans. Through this dataset, we want to emphasize occlusion as a challenging problem for researchers to study, and encourage current algorithms to become more practical for real life situations.

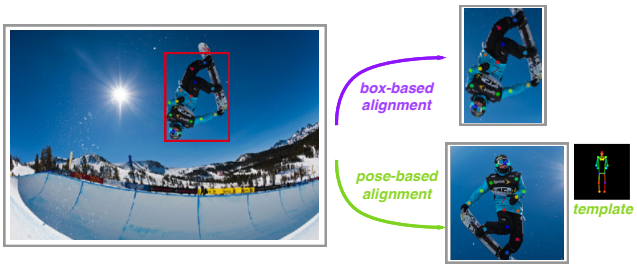

Figure 2: Comparison of box-based alignment and our pose-based alignment (*Affine-Align*). Objects with strange pose are corrected to a standard pose.

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, especially in cases with occlusion.
- We propose a pose-based align module, called *Affine-Align*, which can align image windows into a uniform scale and direction based on human pose.
- We explicitly use artific human *Skeleton* features to guide the segmentation module and achieve a further improvement of the segmentation accuracy.
- We introduce a new benchmark *OCHuman* which focuses on the heavy occlusion problem, with comprehensive annotations including bounding-boxes, human poses and instance masks.

2. Related Work

2.1. Multi-Person Pose Estimation

Top-down methods [2, 6, 14, 16, 29] first employ object detection to crop each person, and then use a single-person pose estimation method on each human instance, which makes them all suffer from the defects of object detection methods on heavy occlusion. While other bottom-up methods [1, 17, 26, 31] first detect body part keypoints of all the people, and then cluster these parts into instances

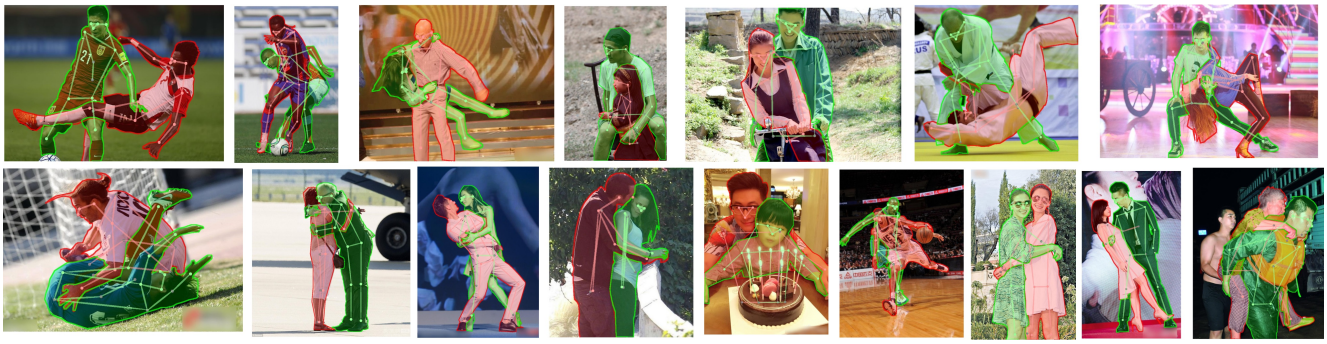

Figure 3: Samples of our *OCHuman* dataset. All the annotated people in this dataset are heavily occluded with others, and have comprehensive annotations.

of human pose. Pishchulin *et al*. [31] 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*. [17] use Resnet [15] to improve precision, and propose image-conditioned pairwise terms to increase speed. Cao *et al*. [1] use knowledge of the human structure, and predict a keypoints heatmap and PAFs, and finally connect the body parts. Newell *et al*. [26] design a tag score map for each body part and use the score map to group body part keypoints.

2.2. Instance Segmentation

Some works [4, 10, 12, 13] employ a multi-stage pipeline which first uses detection to generate bounding boxes and then applies semantic segmentation. Others [3, 19, 22, 30] employ a tighter integration of detection and segmentation, e.g. jointly and simultaneously performing detection and segmentation in an end-to-end framework [19]. Mask R-CNN [14] is the state-of-art performing framework on the COCO [21] dataset competition.

2.3. Harnessing Human Pose Estimation for Instance Segmentation

There are three typical works that combine human pose estimation and instance segmentation. Mask R-CNN [14] approach detects objects while generating instance segmentation and human pose estimation simultaneously in a single framework. But in their work, Mask R-CNN [14] with mask-only performs better than combining keypoints and masks in the instance segmentation task. Pose2Instance [38] proposes a cascade network to harness human pose estimation for instance segmentation. Both of these two works rely on human detection, and perform poorly when two bounding boxes have a large overlap. More recently, PersonLab [28] treats instance segmentation as a pixel-wise clustering problem, and use human pose to refine the clustering results. Although their method is not based on bounding-box detection, they cannot perform as well as Mask R-CNN [14] in the segmentation task.

3. Occluded Human Benchmark

Our *"Occluded Human (OCHuman)"* dataset contains 8110 human instances within 4731 images. Each human instance is heavily occluded by one or several others. We use *MaxIoU* to measure the severity of an object being occluded, which means the max *IoU* with other same category objects in a single image. Those instances with MaxIoU $>$0.5 are referred to as *heavy occlusion*, and are selected to form this dataset. Figure 3 shows some samples from this dataset. With an average of 0.67 MaxIoU for each person, *OCHuman* is the most challenging dataset related to human instances. Moreover, *OCHuman* also has rich annotations. Each instance is annotated with a bounding-box for object detection, an instance binary mask for instance segmentation and 17 body joint locations for pose estimation. All images are collected from real-world scenarios containing people with challenging poses and viewpoints, various appearances and in a wide range of resolutions. With *OCHuman*, we provide a new benchmark for the problem of *occlusion*.

3.1. Annotations

For each image we first annotate the bounding-box of all humans present. Then we calculate the IoU between all the person pairs, and mark those persons with MaxIoU$>$0.5 as heavily occluded instances. Finally, we provide extra information for those occluded instance. The *OCHuman* dataset contains three kinds of annotations related to humans: bounding-boxes, instance binary masks and 17 body joint locations. We reference the definition of body joints from [21], which are eye, nose, ear, shoulder, elbow, wrist, hip, knee and ankle. Except for the nose, all other joints have distinct left and right instances.

3.2. Dataset Splits

OCHuman dataset is designed for validation and testing. Since all the instances in this dataset are heavily occluded by other instances, we consider it is better to use general datasets such as COCO [21] as a training set, then test the robustness of the segmentation methods to occlusion using this dataset, rather than performing training on

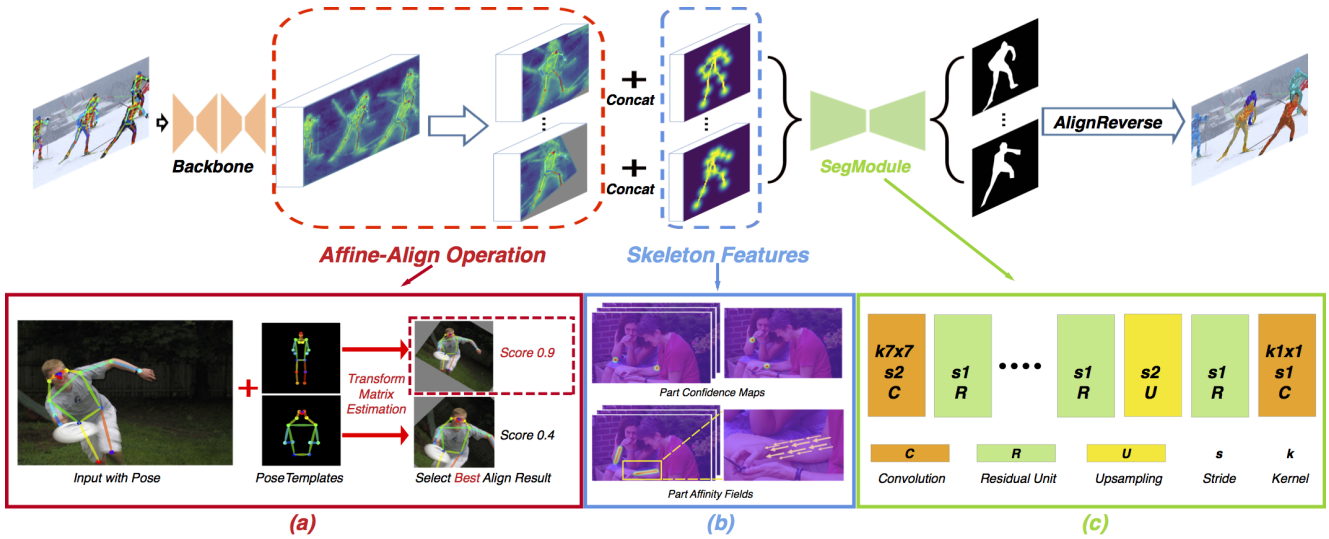

Figure 4: Overview of our network structure (Sec. 4.1). (a) *Affine-Align* operation (Sec. 4.2). (b) Skeleton features (Sec. 4.3). (c) Structure of SegModule (Sec. 4.4), in which residual unit refers to [15].

	COCOPersons (*train+val*)	OCHuman (*val+test*)
#images	**64115**	4731
#persons	**273469**	8110
#persons ($oc_{0.5}$)	2619(<1.0%)	**8110(100%)**
#persons ($oc_{0.75}$)	214(<0.1%)	**2614(32%)**
#average MaxIoU	0.08	**0.67**

Table 1: Comparison of different public datasets related to occluded human. "persons (oc_X)" represents occluded persons with MaxIoU > X.

only occluded cases. We split our dataset into separate validation and test sets. Following random selection, we arrive at a unique split consisting of 2500 validation and 2231 testing image, containing 4313 and 3797 instances respectively. Furthermore, we divide instances in *OCHuman* dataset into two subsets: *OCHuman-Moderate* and *OCHuman-Hard*. The first subset contains instances with MaxIoU in the range of 0.5 and 0.75, while the second contains instances with MaxIoU larger than 0.75, making it the more challenging subset. With these two subsets, we can evaluate the ability of algorithms to handle occlusions of different levels of severity.

3.3. Dataset Statistics

We compare our dataset with the person part of COCO in Table. 1, which is currently the largest public dataset that contains both instance masks and human pose key-points. Although COCO includes comprehensive annotations, it contains few occluded human cases, and so this dataset cannot help to evaluate the capability of methods when faced with occlusion. *OCHuman* is designed for all three most important tasks related to humans: detection, pose estimation and instance segmentation. It is the most challenging benchmark because of its heavy occlusion.

4. Approach

4.1. Overview

Our overall structure is shown in Figure 4, which takes both the image and the human pose as input. Firstly, a base network is used to extract the features of the image. Then an align module, called *Affine-Align*, is used to align *RoIs* to a uniform size, which is 64×64 in this paper, based on the human pose. In the meantime, we generate *Skeleton* features for each human instance and concatenate them to the *RoIs*. Our segmentation module, which we called *SegModule*, is designed based on the same residual unit in Resnet [15]. We carry out experiments on how the depth of *SegModule* contributes to the performance of this system in Section 5.3.3. Finally, we use the estimated matrices in *Affine-Align* operation to reverse the alignment for each instance and get the final segmentation results. We describes our *Affine-Align* operation, *Skeleton* features and *SegModule* in the following subsections.

4.2. Affine-Align Operation

Our *Affine-Align* operation is inspired by the RoI-Pooling in Faster R-CNN [33] and RoI-Align in Mask R-CNN [14]. But unlike them, we align the people based on human pose instead of bounding-boxes. Specifically, as shown in Figure 4(a), we first cluster the poses in the dataset and use the center of each cluster as pose templates, to represent the standard poses in the dataset. Then for each pose detected in the image, we estimate the affine transformation matrix H between it and the templates, and chose the best H based on the transformation error. Finally, we apply H to the image or features and transform it to the desired resolution using bilinear interpolation. Details are introduced below.

4.2.1 Human Pose Representation

Human poses are represented as a list of vectors. Let vector $P = (C_1, C_2, ..., C_m) \in \mathbb{R}^{m\times 3}$ represent the pose of a single person, where $C_i = (x, y, v) \in \mathbb{R}^3$ is a 3D vector representing the coordinates of a single part (such as right-shoulder, left-ankle) and the visibility of this body joint. m is a dataset related parameter meaning the total number of parts in a single pose, which is 17 in COCO.

4.2.2 Pose Templates

We cluster the pose templates from the training set to best represent the distribution of various human poses. We use K-means clustering [7] to cluster the poses $(P_1, P_2, ..., P_n)$ into $k(\leq n)$ sets $\mathbf{S} = \{S_1, S_2, ..., S_k\}$ by optimizing Eq. 1, in which $P_{\mu i}$ is the mean of poses in S_i. We define the distance between two human poses using Eq. 2 and Eq. 3, with several preprocessing steps: (1) We first crop a square-RoI of each instance using its bounding-box, and put the target into the center of the RoI, along with its pose coordinates. (2) We resize this square-RoI to 1×1, so that the pose coordinates are all normalized to $(0, 1)$. (3) We only count those poses which contain more than 8 valid points in the dataset to serve our purpose of creating the pose templates. Poses with few valid points cannot provide effective information and would act as outliers during K-means clustering.

$$\arg\min_{S} \sum_{i=1}^{K} \sum_{P\in S_i} Dist(P, P_{\mu i}) \tag{1}$$

$$Dist(P, P_{\mu i}) = \sum_{j=1}^{m} \|C_j - C_{\mu ij}\|^2 \tag{2}$$

$$C_j = \begin{cases} (x, y, 2) & \text{if } C_j \text{ is visible} \\ (x, y, 1) & \text{if } C_j \text{ is not visible} \\ (0.5, 0.5, 0) & \text{if } C_j \text{ is not in image} \end{cases} \tag{3}$$

After K-means, we use the mean value of each set $P_{\mu i}$ to form the pose template and use it to represent the whole group. We set those body joints with $v > 0.5$ in $P_{\mu i}$ as valid points. Clustering results with different values of K on the COCO training set are shown in Figure 5. Although the results of K-means are heavily reliant on initialization values, our multiple experimental results remain the same, which shows that there is a strong distinction between different sets of human poses. After careful observation of those pose templates, we can find the two most frequent human poses in COCO are a half-body pose and a full-body pose, which is in line with our common sense view of daily life. When $K = 3$ in K-means, we get a half-body pose, a full-body backview and a full-body frontview. When $K \geq 4$, the difference between left and right are introduced. Since our align process copes with the left-right flip, $K \geq 4$ seems unnecessary for our framework. So finally, we choose $K = 3$ to cluster pose templates in our approach.

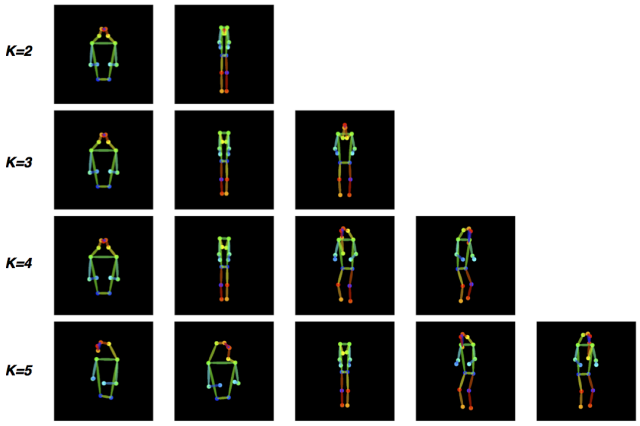

Figure 5: Pose templates clustered using K-means on COCO.

4.2.3 Estimate Affine Transformation Matrix

Let vector P_μ represent a pose template, and P represent a single person pose estimation result. We optimize Eq. 4 to estimate an affine transformation matrix H which transforms the pose coordinates to be as near as possible to the template coordinates. H is a 2×3 matrix with 5 independent variables: rotation, scale factor, x-axis translation, y-axis translation and whether to do left-right flip. Since we have K templates, we define a score for each H^* based on the optimized error value, calculated by Eq. 5, to choose the best template for each estimated pose, as shown in Figure 4(a). In order to get the unique solution from Eq. 4, P_μ and P must contain at least three valid points in common, which can provide at least 6 independent equations for optimizing Eq. 4. If none of our pose templates satisfy this condition, such as the case where there is only one valid point in P, the estimated transformation matrix H^* will be calculated to align the whole image to the desired solution. In most case, this is reasonable because situations lacking valid points in the image mostly correspond to a single, large person in the image.

$$H^* = \arg\min_{H} \|H \cdot P - P_\mu\|. \tag{4}$$

$$score = \exp(-\|H^* \cdot P - P_\mu)\|) \tag{5}$$

4.3. Skeleton Features

Figure 4(b) shows our *Skeleton* features. We adopt the part affinity fields (PAFs) from [1], which is a 2-channel vector field map for each skeleton. We use PAFs to represent the skeleton structure of a human pose. With 19 skeletons defined in the COCO dataset, PAFs is a 38-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. For the COCO dataset, each human pose has a 17-channel part confidence map and a 38-channel PAFs map. So the total number of channels in our *Skeleton* features is 55 for each human instance.

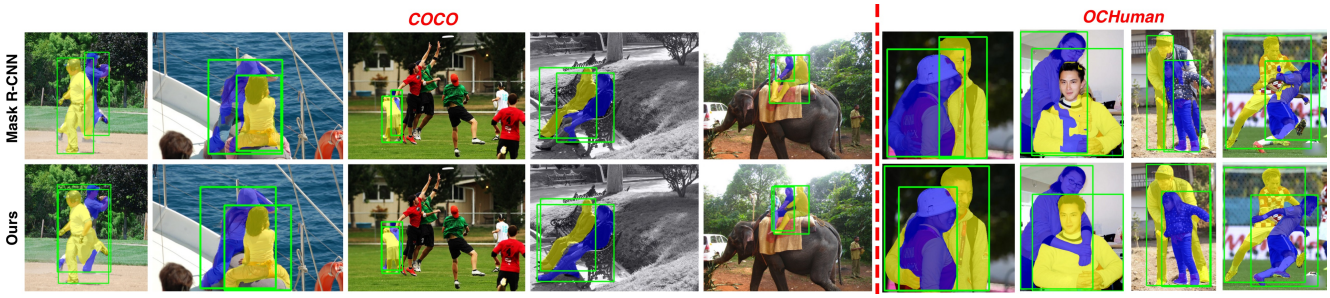

Figure 6: Our method's results vs. Mask R-CNN [14] on occlusion cases. Bounding-boxes in our results are generated using predicted masks for better visualization and comparison.

4.4. SegModule

Since we introduced *Skeleton* features after alignment to artificially extend the image features, Our segmentation module, which we called *SegModule*, needs to have enough receptive fields to not only fully understand these artificial features, but also learn the connections between them and the image features extracted by the base network. Therefore, we design *SegModule* based on the resolution of the aligned RoIs. Figure 4(c) demonstrates the overall architecture of our *SegModule*. It starts with a 7×7, stride-2 convolution layer, and is followed by several standard residual units [15] to achieve a large enough receptive field for the RoIs. After that, a bilinear upsampling layer is used to restore the resolution, and another residual unit, along with a 1×1 convolution layer are used to predict the final result. Such a structure with 10 residual units can achieve about 50 pixels of receptive field, corresponding to our alignment size of 64×64. Fewer units will make the network less capable of learning, and more units enable little improvement on the learning ability. Table 4 shows our experiment on this.

5. Experiments

We evaluate our proposed method on two datasets: (1) *OCHuman*, which is the largest validation dataset that is focused on heavily occluded humans, and proposed in this paper; and (2) *COCOPersons* (the person category of COCO) [21], which contains the most common scenarios in daily life. Note that the *Small* category persons in COCO is not contained in COCOPersons due to the lack of annotations of human pose.

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, so all of our models are trained end-to-end on the COCOPersons training set with the annotations of pose keypoints and segmentation masks. We compare our methods with Mask-RCNN [14], the well known detection based instance segmentation framework. For Mask-RCNN [14], we use the author's released code and configurations from [11], and retrained and evaluate the model on the same dataset as our method. Our framework is implemented using Pytorch. The input resolution of our framework is 512×512 in all experiments. All our models are trained using the same training schedule, which is started by $learningrate = 2e-4$, decayed by 0.1 after 33 epochs, and ended after 40 epochs. Each model is trained on a single TITAN X (Pascal) with $batchsize = 4$ for 80 hours. No special techniques are used, such as iterative training, online hard-case mining, or multi-GPU synchronized batch normalization. Our method with images and keypoints as inputs can run about 20 FPS on a TITAX X (Pascal).

5.1. Performance on occlusion

In this experiment, we evaluate our method's capacity for handling occlusion cases compared with Mask-RCNN [14] on the *OCHuman* dataset. All methods in this experiment are trained on COCOPersons, including our keypoint detector baseline [26] which achieves 0.285 / 0.303 AP on the keypoints task of *OCHuman val / test* set. As shown in Table 2, based on this keypoint detector baseline, our framework can achieve nearly 50% higher than the performance of Mask R-CNN [14] on this dataset. In addition, we test the upper limits of our pose-based framework using ground-truth (GT) keypoints as input, and more than double the accuracy. This demonstrates that with a better keypoint detector our framework can perform far better on occlusion problems. Some results are shown in Figure 6.

5.2. Performance on general cases

In this experiment, we evaluate our model on the *COCOPerson* validation set using groundtruth keypoints as input, and get 0.582 AP on the instance segmentation task. We also evaluate the performance of our model under the predicted pose keypoints using our keypoint detector baseline [26], and achieve 0.555 AP. Mask R-CNN [14] can only achieve 0.532 AP on this same dataset. We further compare our results with a recent work, PersonLab [28]. Scores of PersonLab [28] are taken from their paper, in which the detector is trained and tested on the whole person category of COCO. For fair comparison, we only compare against the results of the *Median* and *Large* categories. Our results surpass theirs with a heavier backbone and multi-scale pre-

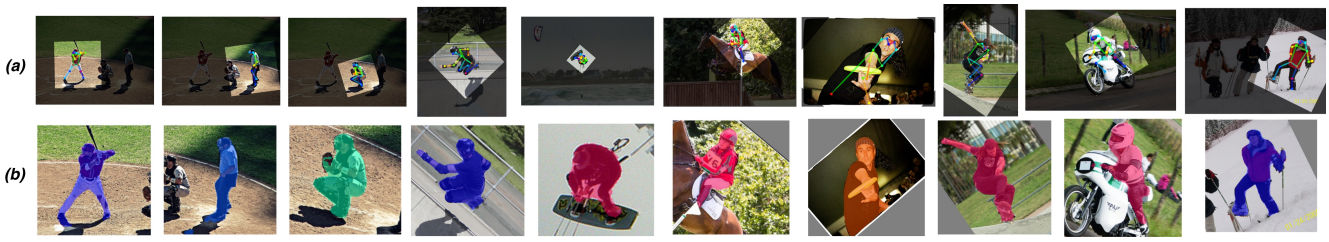

Figure 7: More results of our *Affine-Align* operation. (a) shows the align window on the original image. (b) shows the align results and the segmentation results of our framework.

Methods	*Backbone*	*AP*	AP_M	AP_H
Mask R-CNN	Resnet50-fpn	0.163	0.194	0.113
Ours	Resnet50-fpn	**0.222**	**0.261**	**0.150**
Ours(GT Kpt)	Resnet50-fpn	**0.544**	**0.576**	**0.491**

(a) Performance on OCHuman *val* set.

Methods	*Backbone*	*AP*	AP_M	AP_H
Mask R-CNN	Resnet50-fpn	0.169	0.189	0.128
Ours	Resnet50-fpn	**0.238**	**0.266**	**0.175**
Ours(GT Kpt)	Resnet50-fpn	**0.552**	**0.579**	**0.495**

(b) Performance on OCHuman *test* set.

Table 2: Performance on occlusion. All methods are trained on COCOPersons train split, and tested on *OCHuman*. Ours (GT Kpt) indicates our method with the input of ground-truth keypoints.

Methods	*Backbone*	*AP*	AP_M	AP_L
Mask R-CNN	Resnet50-fpn	0.532	0.433	0.648
PersonLab	Resnet101	-	0.476	0.592
PersonLab	Resnet101(ms scale)	-	0.492	0.621
PersonLab	Resnet152	-	0.483	0.595
PersonLab	Resnet152(ms scale)	-	0.497	0.621
Ours	Resnet50-fpn	**0.555**	**0.498**	**0.670**
Ours(GT Kpt)	Resnet50-fpn	**0.582**	**0.539**	**0.679**

Table 3: Performance on general cases. Mask R-CNN and ours are trained on the COCOPersons train split, and tested on the COCOPersons **val** split (without *Small* category persons). Scores of PersonLab [28] is referred from their paper. Ours (GT Kpt) indicates our method with the input of ground-truth keypoints.

diction, as shown in Table 3. Figure 8 and Figure 7 show some results of our instance segmentation framework and our *Affine-Align* operation, respectively.

5.3. Ablation Experiments

5.3.1 Affine-Align v.s. RoI-Align

Occluded Cases In this experiment, we replace the align module in our framework with RoI-Align based on groundtruth (Gt) bounding-box, and re-train our model with nothing else changed. As shown in Table 5, this box-based alignment strategy can achieve 0.476 AP on *OCHuman* validation set. Our *Affine-Align* based on Gt human pose can achieve 0.544 AP on this same dataset. This means that even if we do not take into account the NMS's deficiencies on handling occlusion (which is eliminated by using GT bounding-boxes), the box-based alignment strategy still does not perform as well as our pose-based alignment strategy in the instance segmentation task of occlusion. The reason is that rotation is allowed in *Affine-Align*, which helps

N Residual Units	5	10	15	20
Receptive Field	~ 30	~ 50	~ 70	~ 90
AP	0.545	0.555	0.555	0.556

Table 4: Experiments on the depth of SegModule under 64×64 RoIs. 10 residual units with a receptive field of about 50 pixels is enough for this alignment size. Deeper architecture brings little benefits. All scores are tested on the COCOPerson *val* set.

to better distinguish two heavy intertwined people by aligning into discriminative RoIs. Strong discrimination RoIs are essential for the segmentation network to locate and extract the specific target.

General Cases We also experiment on COCOPerson validation set. If we allow using both Gt bounding-box and Gt keypoint as input, the best performance is achieved by combining RoI-Align and our *Skeleton* features (0.648 AP). While simultaneously requiring bounding-box and keypoint as input is a rather strict requirement, and both of them can introduce error to the framework when using predicted results instead of ground-truth. If we contraint the framework to only rely on one of them, combining *Affine-Align* with *Skeleton* features can achieved better performance than using RoI-Align stategy on COCOPerson (0.582 AP v.s. 0.568 AP). What's more, the upper limits of the box-based framework is limited by NMS, especially in the case of occlusion. In comparison, our pose-based alignment strategy has no such limits.

Intuitive Pose-based Alignment An intuitive idea of pose-based alignment is to first generate bounding-boxes based on human pose key-points, and then use a box-based alignment strategy, such as RoI-Align, to align each person into a RoI. We take the maximum and minimum values of

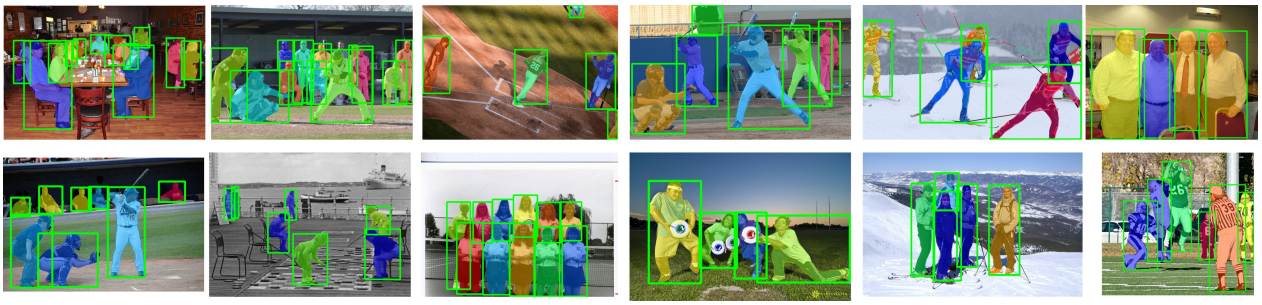

Figure 8: More results of our instance segmentation framework on COCO. Bounding-boxes are generated using predicted masks for better visualization.

the valid key-points as the generated bounding-box, and expand the generated bounding-box by a factor α to simulate the accurate bounding-box as much as possible. We treat α as a hyperparameter and search for the best value during testing. Table 5 shows that no matter how this hyperparameter α is adjusted, the performance still cannot match our *Affine-Align* strategy.

5.3.2 With/Without Skeleton Features

We also experiment on the contribution of our artificial *Skeleton* features. Table 5 shows that our *Skeleton* features are good for different kinds of align strategies because manually concatenating the features of human pose can explicitly provide more information for the network, and lead to a more accurate result. This is more effective for situations where there is more than one person in the RoIs (which is very common), because *Skeleton* features can explicitly guide the network to focus on the specific person. Also, due to this component our framework can better segment the person under occlusion than the previous methods.

5.3.3 SegModule

We have discussed in Section 4.4 that the receptive field is an important factor to be considered in designing the SegModule. So we experiment how the receptive field of SegModule affects our system. We achieve different receptive fields by stacking different numbers of residual units after the first convolution. Besides that, all the other components stay unchanged. As shown in Table 4, our SegModule with 10 residual units can achieve about 50 pixels of receptive field, which is enough for our 64×64 alignment size. A large enough receptive field can provide enough learning ability to understand the image features and artificial features globally. Fewer units will make the network less capable of learning, and more units have little help with the learning ability,

Training Method	Testing Method	BBox Expand	AP *(OCHuman)*	AP *(COCOPerson)*
GT BBOX + RoI-Align (+/-) Skeleton	GT BBOX + RoI-Align (+/-) Skeleton	—	**0.476*/0.133**	**0.648*/0.568**
GT BBOX + RoI-Align (+/-) Skeleton	GT KPT to BBOX + RoI-Align (+/-) Skeleton	30%	0.436/**0.124**	0.431/0.354
		40%	**0.441**/0.115	0.460/0.372
		50%	0.437/0.104	0.477/**0.380**
		60%	0.429/0.093	0.489/0.379
		70%	0.420/0.083	0.497/0.371
		80%	0.411/0.074	**0.501**/0.357
		90%	0.403/0.065	0.500/0.343
		100%	0.393/0.057	0.500/0.325
GT KPT + Affine-Align	GT KPT + Affine-Align	—	**0.544/0.141**	**0.582/0.386**

Table 5: Ablation experiments on OCHuman *val* set and COCOPerson *val* set about different *alignment* strategies and *Skeleton* features. All scores are tested using ground-truth (GT) bounding-box (BBOX) or keypoint (KPT). 'GT KPT to BBOX' represents taking the maximum and minimum values of the valid KPT as the BBOX, and expanding the BBOX by a factor. Notice that scores marked by * rely on both BBOX and KPT as input, while others rely on only one of them.

6. Conclusion

In this paper, we propose a pose-based human instance segmentation framework. We designed an Affine-Align operation for selecting RoIs based on pose instead of bounding-boxes. We explicitly concatenate the human pose skeleton feature to the image feature in the network to further improve the performance. Compared with the traditional detection based instance segmentation framework, our pose-based system can achieve a better performance in the general case, and can moreover better handling occlusion. In addition, we introduce a new dataset called *OCHuman*, which focuses on heavily occluded humans, as a challenging benchmark on occlusion problem.

Acknowledgement: This work was supported by the Natural Science Foundation of China (61772298, 61521002), Research Grant of Beijing Higher Institution Engineering Research Center and Tsinghua-Tencent Joint Laboratory for Internet Innovation Technology.

References

[1] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7291–7299, 2017. 1, 2, 3, 5

[2] Yilun Chen, Zhicheng Wang, Yuxiang Peng, Zhiqiang Zhang, Gang Yu, and Jian Sun. Cascaded pyramid network for multi-person pose estimation. *arXiv preprint arXiv:1711.07319*, 2017. 1, 2

[3] Jifeng Dai, Kaiming He, Yi Li, Shaoqing Ren, and Jian Sun. Instance-sensitive fully convolutional networks. In *European Conference on Computer Vision*, pages 534–549. Springer, 2016. 1, 3

[4] Jifeng Dai, Kaiming He, and Jian Sun. Convolutional feature masking for joint object and stuff segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3992–4000, 2015. 3

[5] Piotr Dollár, Christian Wojek, Bernt Schiele, and Pietro Perona. Pedestrian detection: A benchmark. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 304–311. IEEE, 2009. 2

[6] Hao-Shu Fang, Shuqin Xie, Yu-Wing Tai, and Cewu Lu. RMPE: Regional multi-person pose estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2334–2343, 2017. 1, 2

[7] Edward W Forgy. Cluster analysis of multivariate data: efficiency versus interpretability of classifications. *biometrics*, 21:768–769, 1965. 5

[8] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 3354–3361. IEEE, 2012. 2

[9] Ross Girshick. Fast r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 1440–1448, 2015. 1

[10] Ross Girshick, Forrest Iandola, Trevor Darrell, and Jitendra Malik. Deformable part models are convolutional neural networks. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 437–446, 2015. 3

[11] Ross Girshick, Ilija Radosavovic, Georgia Gkioxari, Piotr Dollár, and Kaiming He. Detectron. https://github.com/facebookresearch/detectron, 2018. 6

[12] Bharath Hariharan, Pablo Arbeláez, Ross Girshick, and Jitendra Malik. Simultaneous detection and segmentation. In *European Conference on Computer Vision*, pages 297–312. Springer, 2014. 3

[13] Bharath Hariharan, Pablo Arbeláez, Ross Girshick, and Jitendra Malik. Hypercolumns for object segmentation and fine-grained localization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 447–456, 2015. 3

[14] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Computer Vision (ICCV), 2017 IEEE International Conference on*, pages 2980–2988. IEEE, 2017. 1, 2, 3, 4, 6

[15] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 3, 4, 6

[16] Shaoli Huang, Mingming Gong, and Dacheng Tao. A coarse-fine network for keypoint localization. In *The IEEE International Conference on Computer Vision (ICCV)*, volume 2, 2017. 2

[17] Eldar Insafutdinov, Leonid Pishchulin, Bjoern Andres, Mykhaylo Andriluka, and Bernt Schiele. DeeperCut: A deeper, stronger, and faster multi-person pose estimation model. In *European Conference on Computer Vision*, pages 34–50. Springer, 2016. 2, 3

[18] Martin Koestinger, Paul Wohlhart, Peter M Roth, and Horst Bischof. Annotated facial landmarks in the wild: A large-scale, real-world database for facial landmark localization. In *2011 IEEE international conference on computer vision workshops (ICCV workshops)*, pages 2144–2151. IEEE, 2011. 1

[19] Yi Li, Haozhi Qi, Jifeng Dai, Xiangyang Ji, and Yichen Wei. Fully convolutional instance-aware semantic segmentation. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 2359–2367, 2017. 1, 3

[20] Masoud Zadghorban Lifkooee, Celong Liu, Yongqing Liang, Yimin Zhu, and Xin Li. Real-time avatar pose transfer and motion generation using locally encoded laplacian offsets. *Journal of Computer Science and Technology*, 34(2):256–271, 2019. 1

[21] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014. 2, 3, 6

[22] Shu Liu, Jiaya Jia, Sanja Fidler, and Raquel Urtasun. Sgn: Sequential grouping networks for instance segmentation. In *The IEEE International Conference on Computer Vision (ICCV)*, 2017. 3

[23] Shuang Liu, Yongqiang Zhang, Xiaosong Yang, Daming Shi, and Jian J Zhang. Robust facial landmark detection and tracking across poses and expressions for in-the-wild monocular video. *Computational Visual Media*, 3(1):33–47, 2017. 1

[24] Xiao Ma, Fandong Zhang, Yuelong Li, and Jufu Feng. Robust sparse representation based face recognition in an adaptive weighted spatial pyramid structure. *Science China Information Sciences*, 61(1):012101, 2018. 1

[25] Jiayuan Mao, Tete Xiao, Yuning Jiang, and Zhimin Cao. What can help pedestrian detection? In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3127–3136, 2017. 1

[26] Alejandro Newell, Zhiao Huang, and Jia Deng. Associative embedding: End-to-end learning for joint detection and grouping. In *Advances in Neural Information Processing Systems*, pages 2277–2287, 2017. 1, 2, 3, 6

[27] Peng Ouyang, Shouyi Yin, Chenchen Deng, Leibo Liu, and Shaojun Wei. A fast face detection architecture for autofocus in smart-phones and digital cameras. *Science China Information Sciences*, 59(12):122402, 2016. 1

[28] George Papandreou, Tyler Zhu, Liang-Chieh Chen, Spyros Gidaris, Jonathan Tompson, and Kevin Murphy. Personlab: Person pose estimation and instance segmentation with

a bottom-up, part-based, geometric embedding model. *arXiv preprint arXiv:1803.08225*, 2018. 3, 6, 7

[29] George Papandreou, Tyler Zhu, Nori Kanazawa, Alexander Toshev, Jonathan Tompson, Chris Bregler, and Kevin Murphy. Towards accurate multi-person pose estimation in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4903–4911, 2017. 2

[30] Pedro O Pinheiro, Ronan Collobert, and Piotr Dollár. Learning to segment object candidates. In *Advances in Neural Information Processing Systems*, pages 1990–1998, 2015. 1, 3

[31] Martin Rajchl, Matthew CH Lee, Ozan Oktay, Konstantinos Kamnitsas, Jonathan Passerat-Palmbach, Wenjia Bai, Mellisa Damodaram, Mary A Rutherford, Joseph V Hajnal, Bernhard Kainz, et al. Deepcut: Object segmentation from bounding box annotations using convolutional neural networks. *IEEE Transactions on Medical Imaging*, 36(2):674–683, 2017. 2

[32] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 779–788, 2016. 1

[33] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*, pages 91–99, 2015. 1, 4

[34] Shuai Shao, Zijian Zhao, Boxun Li, Tete Xiao, Gang Yu, Xiangyu Zhang, and Jian Sun. Crowdhuman: A benchmark for detecting human in a crowd. *arXiv preprint arXiv:1805.00123*, 2018. 2

[35] Xiaoyong Shen, Hongyun Gao, Xin Tao, Chao Zhou, and Jiaya Jia. High-quality correspondence and segmentation estimation for dual-lens smart-phone portraits. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3257–3266, 2017. 1

[36] Xiaoyong Shen, Aaron Hertzmann, Jiaya Jia, Sylvain Paris, Brian Price, Eli Shechtman, and Ian Sachs. Automatic portrait segmentation for image stylization. In *Computer Graphics Forum*, volume 35, pages 93–102. Wiley Online Library, 2016. 1

[37] Xiaoyong Shen, Tao Xin, Hongyun Gao, Zhou Chao, and Jiaya Jia. Deep automatic portrait matting. In *European Conference on Computer Vision*, 2016. 1

[38] Subarna Tripathi, Maxwell Collins, Matthew Brown, and Serge Belongie. Pose2instance: Harnessing keypoints for person instance segmentation. *arXiv preprint arXiv:1704.01152*, 2017. 3

[39] Jie Wang, Juyong Zhang, Changwei Luo, and Falai Chen. Joint head pose and facial landmark regression from depth images. *Computational Visual Media*, 3(3):229–241, 2017. 1

[40] Shihong Xia, Lin Gao, Yu-Kun Lai, Ming-Ze Yuan, and Jinxiang Chai. A survey on human performance capture and animation. *Journal of Computer Science and Technology*, 32(3):536–554, 2017. 1

[41] Liliang Zhang, Liang Lin, Xiaodan Liang, and Kaiming He. Is faster r-cnn doing well for pedestrian detection? In *European conference on computer vision*, pages 443–457. Springer, 2016. 1

[42] Shanshan Zhang, Jian Yang, and Bernt Schiele. Occluded pedestrian detection through guided attention in cnns. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6995–7003, 2018. 1

[43] Zhanpeng Zhang, Ping Luo, Chen Change Loy, and Xiaoou Tang. Facial landmark detection by deep multi-task learning. In *European conference on computer vision*, pages 94–108. Springer, 2014. 1

[44] Erjin Zhou, Haoqiang Fan, Zhimin Cao, Yuning Jiang, and Qi Yin. Extensive facial landmark localization with coarse-to-fine convolutional network cascade. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 386–391, 2013. 1