2020 25th International Conference on Pattern Recognition (ICPR)
Milan, Italy, Jan 10-15, 2021

StrongPose: Bottom-up and Strong Keypoint Heat Map Based Pose Estimation

1st Niaz Ahmad
Department of Computer Science and Engineering
Hanyang University, Korea
niazahmad89@gmail.com

2nd Jongwon Yoon
Department of Computer Science and Engineering
Hanyang University, Korea
jongwon@hanyang.ac.kr

***Abstract*—The adaptation of deep convolutional neural network has made revolutionary advances in human body posture estimation. Various applications utilizing deep neural network for pose estimation have drawn considerable attention in recent years. However, prediction and localization of keypoints in single-person and multi-person images is still a challenging problem. Towards this, we propose a bottom-up approach to pose estimation and motion recognition. We present *StrongPose* system that deals with object-part associations using part-based modeling. The convolution network in our model detects strong keypoint heat maps and predicts their comparative displacements, allowing keypoints to be grouped into human instances. Further, it utilizes the keypoints to generate body heat maps that can determine the position of the human body in the image. The *StrongPose* system is based on fully convolutional engineering and makes proficient inferences while maintaining runtime regardless of the number of individuals in the image. We train and test the *StrongPose* on the COCO dataset. Evaluation results show that our framework achieves average precision of 0.708 using ResNet-101 and 0.725 using ResNet-152. Our results considerably outperform prior bottom-up frameworks.**

***Index Terms*—Body heat map, Pose estimation, Strong keypoint heat map.**

I. Introduction

Recent advances in computer vision have empowered researchers to go beyond classic methods, such as box-level face and body detection, to a comprehensive visual understanding of people in unregulated environments. A deep visual understanding of people is also important in many computer vision applications such as video surveillance, people and activity recognition, human and computer interaction, motion capture and robotics. Human pose estimation is a major cornerstone for achieving a specific goal, described by the two-dimensional positioning of a person's joints, torso and facial keypoints. It is desirable to identify individuals by taking into account the actions they engage in and activities they perform. Recent developments in the estimation of human body have been achieved by improving the efficiency of complex convolutional neural networks (CNNs) and large-scale pose estimation datasets such as MPII [1] and COCO [2].

2D estimation of human posture is a challenging problem. Inferring the poses of multiple people, especially those who are socially engaged, comes with much more difficulty due to the following reasons. First, an image can have an undefined number of individuals that can appear at any location and distance. Second, human-to-human interactions induce complex spatial interference due to contacts, obstruction and articulations of the limbs, making it difficult to associate body parts. Third, the complexity and runtime tend to increase with the number of people in the image, leading to the performance issues.

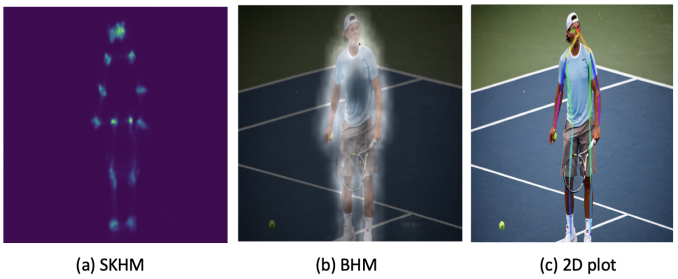

Fig. 1. (a) *StrongPose* generates a strong keypoints heat map (SKHM) for pose detection. (b) Based on the SKHM, a body heat map (BHM) is created to determine the body position. (c) *StrongPose* creates 2D plot with kinematic graph based on the SKHM.

Recent approaches in this domain include a person detector that performs pose estimation for each individual. However, this top-down approach has several problems. The performance of a person detector directly depends on the number of existing individuals in the image. In particular, the larger the crowd, the higher the computing cost, as it requires a single person pose estimator to run iteratively for each detection. When the person detector fails (usually in the case of a crowd), there is no recovery to estimate the pose.

In contrast, a bottom-up approach is attractive because it provides robustness to the problems of a top-down approach and reduces run-time complexity. In practice, methods in [3, 4] fail to maintain the efficiency because the final analysis requires costly global inference. The authors in [3] proposed a method that jointly solves the tasks of detection and pose estimation. However, solving the problem of integral linear programming on a fully connected graph is an NP-hard and takes several hours. Although DeeperCut [4] improved body part detector for effective proposals and group these into valid

This work was supported by Basic Science Research Program through the National Research Foundation of south Korea (NRF) funded by the Ministry of Education NRF-2018R1C1B6006436. Jongwon Yoon is the corresponding author.

human pose configurations, the process time is still in minutes per image. The methods proposed in [5, 6] first predict the human bounding box, then find the keypoints within the box. Specifically, the human keypoints are obtained by transforming the feature map of the detected human bounding box. They require a two-step process of detecting individuals and making poses of detected individuals, leading to double computational power and time. Moreover, addressing the hard keypoints, e.g., torso, knees and ankles, is a difficult problem [7, 8, 9].

We propose *StrongPose* framework to address the difficulties in pose estimation. In particular, *StrongPose* predicts every keypoint of each individual in a completely convolutional manner. Our system introduces strong keypoint heat map (SKHM, shown in Figure 1(a)) scheme that estimates the relative displacement between each pair of keypoints and significantly improves the precision of long range, occluded and close proximity keypoints. Once the keypoints are localized, a fast pose plotting algorithm is used to organize them into instances. Our method starts with the most obvious identification, as opposed to starting with the dominant body parts such as the nose, therefore it always work well even with debris or unclear visibility. Our framework can successfully process the pose estimation of both single-person and multi-person scenarios.

We evaluate the performance of *StrongPose* on the COCO keypoint dataset annotated by multiple persons with 5 facial and 12 body keypoints. *StrongPose* surpasses the best prior bottom-up technique [7] by increasing the average precision from 0.696 to 0.725. In addition, producing body heat map (BHM, shown in Figure 1(b)) help to localize the human body and improves the keypoint confidence in the scene. Our algorithm is very fast and simple because it does not require two-step box-based refinement nor clustering. With this reason, we believe *StrongPose* is quite useful for various applications, such as AR/VR, video surveillance, action recognition and many others. Our contributions are multi-fold:

- We propose a novel and efficient approach *StrongPose*, which generate SKHM for both soft and hard keypoints.
- Using SKHM, *StrongPose* generates BHM in order to localize the individuals in the image.
- We investigate the effects of various factors that contribute to single and multi-person pose estimation in bottom-up approach.
- The performance of *StrongPose* on challenging COCO keypoint benchmark is 0.708 average precision (AP) using ResNet-101 and 0.725 AP using ResNet-152.

The rest of this paper is organized as follows. We discuss related works on pose estimation in Section II. Section III describes our *StrongPose* system and key algorithm for pose estimation. We present implementation details and evaluation results in Section IV. Section V concludes the paper.

II. Related work

Early human posture prediction relies on the inference mechanism of part-dependent graphical model where humans are defined by a set of configurable parts [10, 11]. In these years, the trend toward deep convolutional neural networks (CNN) is increasing [12, 13]. In the deep CNN domain, [14, 15, 16] proposed tractable inference algorithms to solve the energy minimization of abundant dependency among body parts. Model-based large-scale convolutional networks have achieved state-of-the art performance in both single-person and multi person-pose estimation. In *StrongPose*, the forward inference method differs from these early deformable part-based models; we follow a bottom-up methodology for individual keypoints detection and group them into individual instances.

Mainly, there are two approaches for estimating human posture; bottom-up (parts first) and top-down (person first). The top-down approach identifies keypoints surrounded by bounding box detector. The bottom-up method first detects all human keypoints (e.g., ears, eyes, nose and joints) in the whole image, and then groups these keypoints into human instance.

Examples of bottom-up approach are [4] and [17], both are based on ResNet [18]. The algorithms generate powerful part detectors and the image is dependent on pairwise scores, which improves the process and runtime. However, these methods take time in minutes to perform pose estimation for each image if the number of proposals is small. Pishchulin *et al*. [3] formulate a multi-person pose estimation problem with labeling and part-grouping through a linear program on a fully connected graph that is an NP-hard problem and requires significant computational power. The authors in [7] propose an associative embedding for the identification of keypoint detections from the same individual. Coa *et al*. [19] map associations between keypoints into Part Affinity Fields (PAF) and groups candidate joints into a person using greedy algorithm.

In general, a top-down approach predicts the locations of keypoints inside boundary boxes obtained from a person detector (e.g., Fast-RCNN [20], Faster-RCNN [21] and R-FCN [22]). Mask R-CNN [5] extends the Faster R-CNN [21] by adding a predicting segmentation masks to each region of interest, however its processing time is 5 frames per second, which is not compatible with real-time scenario. Iqbal *et al*. [23] address the joint-to-person association problem using a densely connected graphical model. They first crop image regions using a person detector, then process the cropped area with integer linear programming. Their algorithm is based on HOG system that relies heavily on training data, however it does not perform well at handling various pose estimations. Chen *et al*. [6] propose a two-level Cascaded Pyramid Network (CPN) consists of GlobalNet and RefineNet to address the hard-joint problem. The downside of hard-keypoints is that they require more context than nearby features. Fang *et al*. [24] propose a framework, Symmetric Spatial Transformer Network (SSTN), to extract high quality single-person region from inaccurate bounding boxes. The SSTN depends on the concept of a bounding box that requires processing time for pose estimation. The authors in [25] address the robustness problem of keypoint localization using a fully convolutional inception network [26] and build both coarse-part and fine-part detectors for feature extraction. The coarse detector can yield accurate location of keypoints with distinctive shape, however

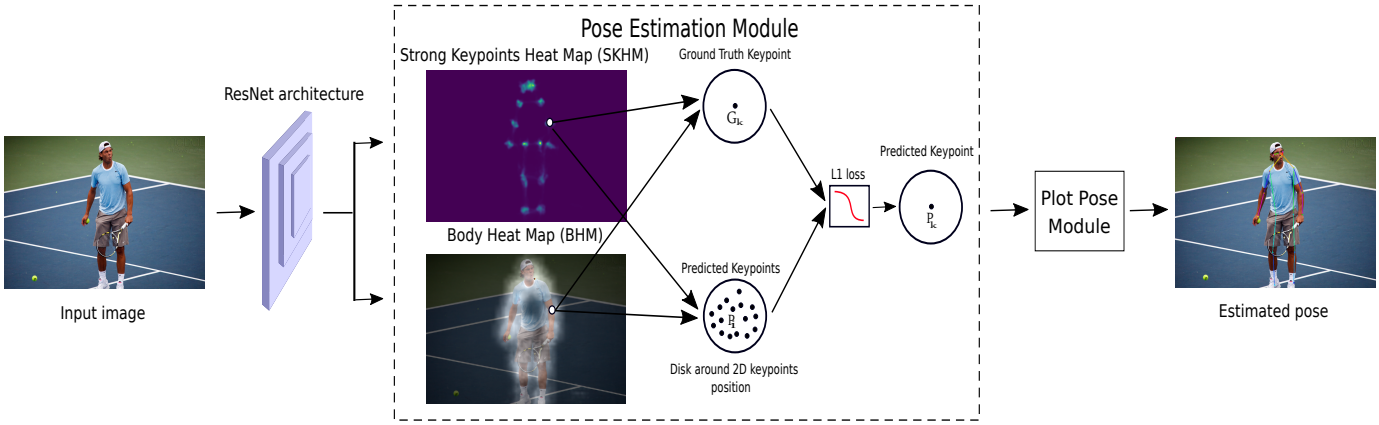

Fig. 2. The CNN model in *StrongPose* predicts two heat maps, strong keypoints heat maps (SKHM) and body heat map (BHM). The first prediction identifies keypoints to detect a person's poses, and the second prediction generates the human body heat map to localize each person in the picture.

often fail to detect keypoints with ambiguous appearance. Li *et al*. [27] propose a Multi-Stage Pose estimation Network (MSPN) adopting the GlobalNet of CPN [6]. GlobalNet provides high spatial resolution for localization but low semantic information for recognition.

III. System Overview

Figure 2 depicts an overview of the *StrongPose* system which consists of the ResNet, a pose estimation module and a pose plotting module. The CNN model takes a picture as an input and passes it to the backbone network ResNet. It generates keypoint proposals for each labeled keypoint. These keypoint proposals help with keypoint localization and are fetched into the pose estimation module. The pose estimation module then generates strong keypoint heat maps (SKHM) based on the keypoint proposals and combines the SKHMs. The pose estimation module also produces a body heat map (BHM) utilizing the SKHM in order to identify the position of each individual in the scene. We avoid bounding box level detection due to the BHM. When both the SKHM and the BHM are obtained, the pose estimation module predicts each individual's keypoints and use the L1 loss function to minimize the loss between the prediction and the ground truth. The output of the pose estimation module feeds into the pose plotting module (inspired by [9]) to create a perfect pose for each person in the image.

A. Bottom-Up Approach used in StrongPose

The current-state-of-art of human pose recognition and estimation can be divided into two classes; the bottom-up and top-down approach. The bottom-up approach first finds all person keypoints (e.g. eyes, ears, joints and etc.) in the entire image and then groups them into human instances to create a kinematic graph (human tree-structure). In contrast, top-down method starts with human detection and identifies keypoints within the detection bounding boxes. This is based on the assumption that each bounding box contains at most one keypoint per keypoint class.

Fig. 3. Example of strong keypoints heat map generated by *StrongPose*.

In *StrongPose*, we adopt the bottom-up approach, which is more suitable for crowded scenario, because it processes the entire image at once and its runtime is independent of the number of individuals in the image. *StrongPose* first detects body parts and then groups these parts into human instances. In this manner, *StrongPose* is able to reduce per image processing time, save computational power for training and achieve confident result in crowded scene.

B. Strong Keypoint

Strong keypoint detection is the core engine of *StrongPose*. In an instance-agnostic manner, it detects all visible keypoint proposals that belong to every individuals in the image. In order to detect all keypoint proposals, we significantly modified the hybrid classification and regression approach used in [28] and adopted it into the multi-person framework. Figure 3 presents an example of the SKHM generated by *StrongPose*.

The system generates the SKHM using keypoint offsets, two channels per keypoint for vertical and horizontal displacement. Suppose p_i is the position of the 2D keypoint in the image, where $i = 1, \ldots, N$ is the indexing position of the image and N is the number of pixels. Let $D_R(q) = \{p : \|p–q\| \leq R\}$ is a disk of radius R centered on q. Let q_{jk} is the 2D position of the k-th keypoint of the j-th person instance (with $j = 1, \ldots, I$, where I is the number of person instances in the image). For each keypoint $k = 1, ..., K$, a binary classification task is set

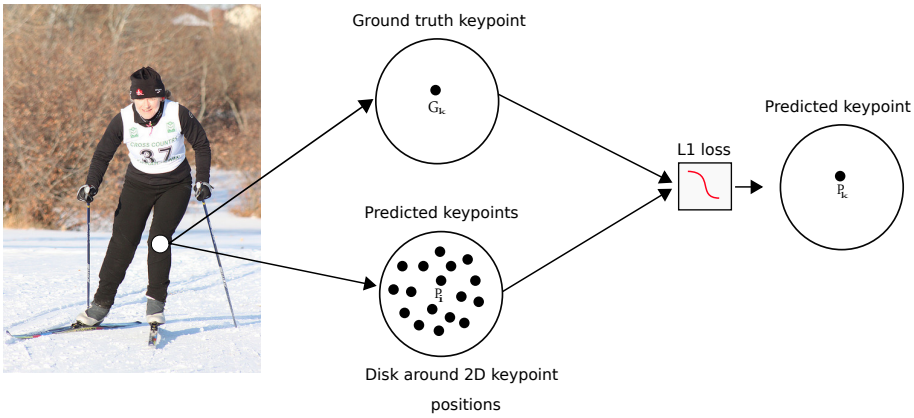

Fig. 4. Keypoint disk around the keypoint positions. L1 loss function is used to minimize the error, which is the sum of all absolute differences between the ground truth and predicted keypoint values.

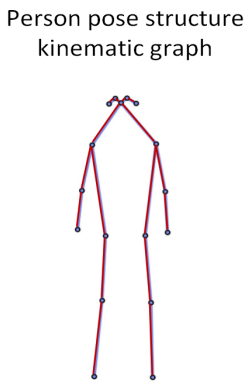

Fig. 5. *StrongPose* defines keypoints and body parts association.

as follows. The predicted keypoint heat map $p_k(h) = 1$ if $h \in D_R$ for every person instance j, otherwise $p_k(h) = 0$. Thus, for each keypoint type we have K independent dense binary classification tasks. The radius value is set to R = 16 pixels to predict a disk of radius R around a particular keypoint of any person in the image (we empirically obtain the R value and set it to 16 which reaches nearest to the ground truth). The value R is constant for all experiments reported in this paper. In order to equally weigh all person instances in the classification, we choose a disk radius which does not scale according to the instance size. While training the network, the keypoint heat map loss is computed as the average logistic loss based on the image position. It then back-propagates across the entire image, excluding the range that includes individuals who are not fully annotated with keypoints (e.g., crowded areas and small individual segments).

In expansion to the heat maps, *StrongPose* also predicts the vector of keypoint offsets $V_k(x)$ which is used to increase the precision of keypoint localization. The keypoint offset 2D vector $V_k(x) = q_{jk} - p$ for each position p_i inside the keypoint disk and for each keypoint type k. It focuses the picture position p to the k-th keypoint of the closest individual instance j. It also generates vector fields V while solving the 2D regression problem at each picture position and keypoint. During training, the prediction error of the keypoint offset is penalized by the L1 loss. As presented in Figure 4, the error is averaged and back-propagated only at the positions $p \in D_R$, where the ground truth keypoint position is G_k. We reduce the errors in the keypoint offset (radius R = 16 pixels) by normalizing them and making a dynamic range compatible with the heat map classification loss.

C. Pose Plotting

StrongPose utilizes a pose plotting algorithm to collect all keypoints and turn them into individual instances. Initially, a queue stores all *K* keypoints along with the keypoint position x_i and type k. However, two or more keypoints can be selected for one keypoint (x and y coordinates). For all these local maxima (two or more keypoints) within the keypoint heat map $h_k(x)$, a Gaussian filter with a threshold value of 1.0 is used to select the keypoint with high intensity. These points are used as building blocks to detect instances. It then gradually connects adjacent pairs of keypoints along the edges of the kinematic graph depicted in Figure 5. At each iteration, if the location x_i of the current detection point of type k is inside a disk $D_R(q_{j',k})$ of instance j', then algorithm skips such point because it is already recognized. It usually occurs when two keypoints overlap or partially touch. For this matter, we utilize a non-maximum suppression. Then a new detection instance j starts with the k-th keypoint at location $q_{jk} = x_i$ and delivers it to the new point.

Other approaches of plotting poses based on the torso or nose keypoints sometimes fail to plot individuals when they are not clearly visible in the image. Although our decoding (and plotting) algorithm does not take into account both kinds of keypoints, we found that *StrongPose* easily identifies the frontal face of an individual (from facial keypoints). Moreover, it manages rigorous situations where a significant part of the individual is not visible.

D. Confidence Score

We have experimented with various strategies for assigning keypoint confidence scores to detections generated by our fast pose plotting algorithm. We adopted the method used in [28] and assigned a confidence score $s_{jk} = h_k(q_{jk})$ to each key point, however there was a fairness issue in our configurations. Specifically, well-localized facial keypoints generally obtain significantly higher score than poorly positioned keypoints such as hip or knee. Therefore, we calculate the score of various kinds of keypoints inspired by the object keypoint similarity (OKS) evaluation metric used in COCO. We use precision threshold T_k to penalize localization errors at different types of keypoints.

For keypoint confidence score, we adopt non-maximum suppression (NMS) and utilize the average of the keypoint scores as the instance level score $s_j^h = (1/k) \sum_k s_{jk}$. The

tests are conducted on both hard OKS-based NMS [28] and soft-NMS schemes [29]. We use the sum of the keypoint scores, not claimed by the higher scored instances, as the final instance-level score. It is also normalized to the total number of keypoints as follows:

$$s_j = (1/k) \sum_{k=1:K} s_{jk}[||q_{jk} - q_{j'k}|| > r, \forall j' < j] \quad (1)$$

where the NMS-radius r = 10. Experimental results are presented with the best Expected-OKS and soft-NMS scores in Section IV.

E. Undefined Keypoint Annotation

The COCO dataset does not provide keypoint annotations for small instances of the training dataset, and hence our model avoids them during evaluation. However, for such small instances, our system includes the segmentation annotation and evaluates the mask prediction. Since keypoint annotations are required for training the model, we crop around the ground truth box annotations of these small-person instances to define the missing keypoint annotations, and then execute the single-person pose estimator. We treat these keypoints as systematic training annotations while training the *StrongPose* model. The keypoint annotations mentioned above are particularly important for *StrongPose*'s performance on small instances. Note that unlike [30], we do not use any data in this process other than the split images and annotations from the COCO train dataset. The performance of the *StrongPose* can be further improved by purifying the data of additional images as described in [30].

TABLE I
PERFORMANCE COMPARISON WITH HOURGLASS [33], CPN [6], HRNET-W48 [31], CMU-POSE [19] AND PERSONLAB [9] ON COCO VAL2017 DATASET (COMPARISON RESULTS ARE CITED FROM [6, 31]).

Method	Backbone	Input Size	OHKM	AP	AR
Top-down:					
8-stage Hourglass	-	256 × 192	✗	0.669	-
8-stage Hourglass	-	256 × 256	✗	0.671	-
CPN	ResNet-50	256 × 192	✗	0.686	-
CPN	ResNet-50	384 × 288	✗	0.706	-
CPN	ResNet-50	256 × 192	✓	0.694	-
CPN	ResNet-50	384 × 288	✓	0.716	-
HRNet-W48	HRNet-W48	384 × 288	✗	0.763	0.812
Bottom-up:					
CMU-Pose	-	-	✗	0.618	-
PersonLab (single-scale)	ResNet-152	-	✗	0.665	0.707
PersonLab (multi-scale)	ResNet-152	-	✗	0.687	-
StrongPose	ResNet-101	-	✗	0.690	0.757
StrongPose	ResNet-152	-	✗	0.728	0.800

IV. EXPERIMENTS AND RESULTS

A. Dataset and Evaluation Metric

We assess the performance of the *StrongPose* framework on the standard COCO keypoint dataset (person class alone). The COCO keypoint contains the challenge of localizing multi-person keypoint in complex uncontrolled environment. The COCO training, validation, and testing dataset contains over 200K images and 250K human instances labeled with keypoints. In addition, 150K keypoints are open to the public for training and evaluation. Our model is trained only on the COCO train 2017 dataset which includes 57K images and 150K instances of people without any additional data. The ablation is studied on the COCO val2017 dataset. For a fair comparison, we present the results for the test-dev2017 set using a bottom-up approach with the state-of-the-art results [7], [8], [9], [19]. The COCO evaluation determines the object keypoint similarity (OKS) and uses both the average precision (AP) and average recall (AR) for 10 OKS thresholds as the evaluation metric [32]. The performance is determined from the difference between the predicted point and the ground truth, normalized to the person's scale.

TABLE II
THE PERFORMANCE OF AP ON THE COCO KEYPOINT TEST-DEV SPLIT. AP AT IOU=.5:.05:.95, $AP^{\cdot 50}$ AT IOU=.50 (PASCAL VOC METRIC), $AP^{\cdot 75}$ AT IOU=.75 (STRICT METRIC), AP^M CORRESPONDS TO AP FOR MEDIUM OBJECTS: $32^2 <$ AREA $< 96^2$, AND AP^L CORRESPONDS TO AP FOR LARGE OBJECTS: AREA $> 96^2$.

Method	AP	$AP^{\cdot 50}$	$AP^{\cdot 75}$	AP^M	AP^L
Top-down:					
Mask-RCNN [5]	0.631	0.873	0.687	0.578	0.714
G-RMI COCO-only [28]	0.649	0.855	0.713	0.623	0.700
CPN [6]	0.721	0.914	0.800	0.687	0.772
Bottom-up:					
CMU-Pose [19] (+refine)	0.618	0.849	0.675	0.571	0.682
Assoc. Embed. [7] (multi-scale)	0.630	0.857	0.689	0.580	0.704
Assoc. Embed. [7] (mscale, refine)	0.655	0.879	0.777	0.690	0.752
PersonLab [9] (single-scale)	0.665	0.880	0.726	0.624	0.723
PersonLab [9] (multi-scale)	0.687	0.890	0.754	0.641	0.755
MultiPoseNet [8]	0.696	0.863	0.766	0.650	0.763
StrongPose:					
ResNet101	0.708	0.889	0.752	0.652	0.753
ResNet152	0.725	0.891	0.778	0.671	0.762

TABLE III
THE PERFORMANCE OF AR ON THE COCO KEYPOINTS TEST-DEV SPLIT.

Method	AR	$AR^{\cdot 50}$	$AR^{\cdot 75}$	AR^M	AR^L
Top-down:					
Mask-RCNN [5]	0.697	0.916	0.749	0.637	0.778
G-RMI COCO-only [28]	0.697	0.887	0.755	0.644	0.771
CPN [6]	0.785	-	-	-	-
Bottom-up:					
CMU-Pose [19] (+refine)	0.665	0.872	0.718	0.606	0.746
Assoc. Embed. [7] (multi-scale)	-	-	-	-	-
Assoc. Embed. [7] (mscale, refine)	0.758	0.912	0.819	0.714	0.820
PersonLab [9] (single-scale)	0.710	0.903	0.766	0.661	0.777
PersonLab [9] (multi-scale)	0.754	0.927	0.812	0.697	0.830
MultiPoseNet [8]	0.735	-	-	-	-
StrongPose:					
ResNet101	0.721	0.904	0.782	0.670	0.771
ResNet152	0.751	0.919	0.802	0.690	0.813

B. Training Model

We use CNN backbone models ResNet-101 and ResNet-152 [18] for training and testing. The ResNet backbone network is initialized with pre-training on ImageNet classification task [34]. We set the learning rate for training to $1 \times e^{-3}$. An image size of 401×401 is fed into the network, while the batch size of two images is processed on NVIDIA GeForce GTX 1080 Ti (GPU is powerful enough to process multiple images at once). We conduct synchronous training for 5000 epochs with stochastic gradient descent, momentum value set to 0.9, and Polyak-Ruppert model parameter averaging. We fix the ResNet activation statistics to ImageNet values with batch normalization [35]. Our ResNet CNN backbone network uses an output stride of 16 during training and reduces to 8 during

Fig. 6. Visualization of COCO val and test images.

evaluation using atrous convolution [36]. To speed up training, we make model predictions during training using the feature activation from a layer in the middle of the network, which we found empirically. For evaluation, we present results of two models trained on ResNet-101 and ResNet-152. The system uses the TensorFlow [37] platform for implementation and all results are obtained from the same model. The algorithm of *StrongPose* is open to the public [38].

C. COCO Keypoints Evaluation

Table I shows the results on COCO val2017 dataset compared with 8-stage Hourglass [33], CPN [6], HRNet-W48 [31], CMU-Pose [19] and PersonLab [9]. The first three methods use a top-down approach for estimating keypoints, while the last two use a bottom-up approach. *StrongPose* increases average precision (AP) by 0.057 compared to Hourglass. Both methods have no Online Hard Keypoints Mining (OHKM) involved. We can see that *StrongPose* outperforms CPN by AP of 0.022 when OHKM is not used and by AP of 0.012 when OHKM is used. In addition, our model provides significant performance improvement over the bottom-up method, CMU-Pose and PersonLab. Specifically, *StrongPose* improves AP of 0.063 and AR of 0.093 compared to PersonLab.

Table II and Table III show the performance of AP and AR

on COCO keypoint test-dev2017 dataset, respectively. We can see that *StrongPose* outperforms bottom-up approaches, CMU-Pose, Associative Embedding, PersonLab and MultiPoseNet. Specifically, our best result yields 0.725 AP on the ResNet-152 base architecture. Some of the bottom-up approaches mentioned above perform multi-scale inference and optimize their results using a single-person pose estimation (adding the results on top of their bottom-up identification proposals). The test results also show that *StrongPose* surpasses the performance of top-down approaches such as Mask-RCNN and G-RMI.

In Figure 6, we present the visualization results generated by *StrongPose* using COCO val and test images. The first two rows show the results of the single-person pose estimation (uncrowded scenarios) and the last two rows show the results of the multi-person pose estimation (crowded scenarios). We can confirm that *StrongPose* accurately predicts the human posture regardless of the number of individuals in the image.

V. Conclusion

We proposed a bottom-up approach using unified part-based modeling to jointly solve pose estimation and person detection problem. *StrongPose* generates both strong keypoints heat maps and body heat map to accurately predict the keypoints. The effectiveness of *StrongPose* is evaluated on the COCO 2017 keypoint challenging dataset and shows cutting-edge results. We believe that *StrongPose* model is simple yet effective architecture for identifying human actions. We can also improve our model while enhancing the algorithm performance and training on multiple datasets. Take the advantages of *StrongPose* model, we plan to develop an autonomous system that recognizes human actions in real-time scenarios in future work.

References

[1] M. Andriluka, L. Pishchulin, P. Gehler, and B. Schiele, "2D Human Pose Estimation: New Benchmark and State of the Art Analysis," in *Proceedings of the IEEE Conference on computer Vision and Pattern Recognition*, pp. 3686–3693, 2014.

[2] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft COCO: Common Objects in Context," in *European Conference on Computer Vision*, pp. 740–755, 2014.

[3] L. Pishchulin, E. Insafutdinov, S. Tang, B. Andres, M. Andriluka, P. V. Gehler, and B. Schiele, "Deepcut: Joint Subset Partition and Labeling for Multi Person Pose Estimation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4929–4937, 2016.

[4] E. Insafutdinov, L. Pishchulin, B. Andres, M. Andriluka, and B. Schiele, "Deepercut: A Deeper, Stronger, and Faster Multi-Person Pose Estimation Model," in *European Conference on Computer Vision*, pp. 34–50, 2016.

[5] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask R-CNN," in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 2961–2969, 2017.

[6] Y. Chen, Z. Wang, Y. Peng, Z. Zhang, G. Yu, and J. Sun, "Cascaded Pyramid Network for Multi-Person Pose Estimation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 7103–7112, 2018.

[7] A. Newell, Z. Huang, and J. Deng, "Associative Embedding: End-to-End Learning for Joint Detection and Grouping," in *Advances in Neural Information Processing Systems*, pp. 2277–2287, 2017.

[8] M. Kocabas, S. Karagoz, and E. Akbas, "Multiposenet: Fast Multi-Person Pose Estimation Using Pose Residual Network," in *Proceedings of the European Conference on Computer Vision*, pp. 417–433, 2018.

[9] G. Papandreou, T. Zhu, L.-C. Chen, S. Gidaris, J. Tompson, and K. Murphy, "PersonLab: Person Pose Estimation and Instance Segmentation With a Bottom-up, Part-based, Geometric Embedding Model," in *Proceedings of the European Conference on Computer Vision*, pp. 269–286, 2018.

[10] M. A. Fischler and R. A. Elschlager, "The Representation and Matching of Pictorial Structures," in *IEEE Transactions on Computers*, vol. 100, no. 1, pp. 67–92, 1973.

[11] P. Felzenszwalb, D. McAllester, and D. Ramanan, "A Discriminatively Trained, Multiscale, Deformable Part Model," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1–8, 2008.

[12] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-Based Learning Applied to Document Recognition," in *IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.

[13] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet Classification With Deep Convolutional Neural Networks," in *Advances in Neural Information Processing Systems*, pp. 1097–1105, 2012.

[14] S. Johnson and M. Everingham, "Learning Effective Human Pose Estimation From Inaccurate Annotation," in *CVPR*, pp. 1465–1472, 2011.

[15] L. Pishchulin, M. Andriluka, P. Gehler, and B. Schiele, "Poselet Conditioned Pictorial Structures," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 588–595, 2013.

[16] B. Sapp and B. Taskar, "Modec: Multimodal Decomposable Models for Human Pose Estimation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3674–3681, 2013.

[17] E. Insafutdinov, M. Andriluka, L. Pishchulin, S. Tang, E. Levinkov, B. Andres, and B. Schiele, "Arttrack: Articulated Multi-Person Tracking in the Wild," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 6457–6465, 2017.

[18] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770–778, 2016.

[19] Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh, "Realtime Multi-Person 2D Pose Estimation Using Part Affinity Fields," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 7291–7299, 2017.

[20] R. Girshick, "Fast R-CNN," in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 1440–1448, 2015.

[21] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards Real-Time Object Detection With Region Proposal Networks," in *Advances in Neural Information Processing Systems*, pp. 91–99, 2015.

[22] J. Dai, Y. Li, K. He, and J. Sun, "R-FCN: Object Detection via Region-Based Fully Convolutional Networks," in *Advances in Neural Information Processing Systems*, pp. 379–387, 2016.

[23] U. Iqbal and J. Gall, "Multi-Person Pose Estimation With Local Joint-to-Person Associations," in *European Conference on Computer Vision*, pp. 627–642, 2016.

[24] H.-S. Fang, S. Xie, Y.-W. Tai, and C. Lu, "RMPE: Regional Multi-Person Pose Estimation," in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 2334–2343, 2017.

[25] S. Huang, M. Gong, and D. Tao, "A Coarse-Fine Network for Keypoint Localization," in *Proceedings of the IEEE Interna-*

tional Conference on Computer Vision, pp. 3028–3037, 2017.

[26] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going Deeper With Convolutions," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1–9, 2015.

[27] W. Li, Z. Wang, B. Yin, Q. Peng, Y. Du, T. Xiao, G. Yu, H. Lu, Y. Wei, and J. Sun, "Rethinking on Multi-Stage Networks for Human Pose Estimation," *arXiv preprint arXiv:1901.00148*, 2019.

[28] G. Papandreou, T. Zhu, N. Kanazawa, A. Toshev, J. Tompson, C. Bregler, and K. Murphy, "Towards Accurate Multi-Person Pose Estimation in the Wild," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4903–4911, 2017.

[29] N. Bodla, B. Singh, R. Chellappa, and L. S. Davis, "Soft-NMS–Improving Object Detection With One Line of Code," in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 5561–5569, 2017.

[30] I. Radosavovic, P. Dollár, R. Girshick, G. Gkioxari, and K. He, "Data Distillation: Towards Omni-Supervised Learning," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4119–4128, 2018.

[31] K. Sun, B. Xiao, D. Liu, and J. Wang, "Deep High-Resolution Representation Learning for Human Pose Estimation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5693–5703, 2019.

[32] Tsung-Yi Lin, Genevieve Patterson, Matteo R. Ronchi, Yin Cui, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, Larry Zitnick, Piotr Dollár , "Coco," http://cocodataset.org, 2015.

[33] A. Newell, K. Yang, and J. Deng, "Stacked Hourglass Networks for Human Pose Estimation," in *European Conference on Computer Vision*, pp. 483–499, 2016.

[34] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, and M. Bernstein, "Imagenet Large Scale vVsual Recognition Challenge," in *International Journal of Computer Vision*, vol. 115, no. 3, pp. 211–252, 2015.

[35] S. Ioffe and C. Szegedy, "Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift," *arXiv preprint arXiv:1502.03167*, 2015.

[36] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs," in *IEEE transactions on Pattern Analysis and Machine Intelligence*, vol. 40, no. 4, pp. 834–848, 2017.

[37] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S. Corrado, A. Davis, J. Dean, and M. Devin, "Tensorflow: Large-Scale Machine Learning on Heterogeneous Distributed Systems," *arXiv preprint arXiv:1603.04467*, 2016.

[38] N. Ahmad and J. Yoon, "StrongPose", https://github.com/niazahamd89/StrongPose/